Natural selection and genetic diversity in the butterfly heliconius melpomene

Natural Selection and Genetic Diversity in the Butterfly *Heliconius melpomene*

Simon H. Martin,*,[1] Markus Möst,* William J. Palmer,[†] Camilo Salazar,[‡] W. Owen McMillan,[§] Francis M. Jiggins,[†] and Chris D. Jiggins*

*Department of Zoology and [†]Department of Genetics, University of Cambridge, CB2 3EH, United Kingdom, [‡]Biology Program, Faculty of Natural Sciences and Mathematics, Universidad del Rosario, Bogota 111221, Colombia, and [§]Smithsonian Tropical Research Institution, Apartado 0843–03092, Balboa, Ancón, Panama

ABSTRACT A combination of selective and neutral evolutionary forces shape patterns of genetic diversity in nature. Among the insects, most previous analyses of the roles of drift and selection in shaping variation across the genome have focused on the genus *Drosophila*. A more complete understanding of these forces will come from analyzing other taxa that differ in population demography and other aspects of biology. We have analyzed diversity and signatures of selection in the neotropical *Heliconius* butterflies using resequenced genomes from 58 wild-caught individuals of *Heliconius melpomene* and another 21 resequenced genomes representing 11 related species. By comparing intraspecific diversity and interspecific divergence, we estimate that 31% of amino acid substitutions between *Heliconius* species are adaptive. Diversity at putatively neutral sites is negatively correlated with the local density of coding sites as well as nonsynonymous substitutions and positively correlated with recombination rate, indicating widespread linked selection. This process also manifests in significantly reduced diversity on longer chromosomes, consistent with lower recombination rates. Although hitchhiking around beneficial nonsynonymous mutations has significantly shaped genetic variation in *H. melpomene*, evidence for strong selective sweeps is limited overall. We did however identify two regions where distinct haplotypes have swept in different populations, leading to increased population differentiation. On the whole, our study suggests that positive selection is less pervasive in these butterflies as compared to fruit flies, a fact that curiously results in very similar levels of neutral diversity in these very different insects.

KEYWORDS background selection; genetic hitchhiking; recombination rate; selective sweeps; effective population size

GENETIC variation within and between populations is shaped by numerous factors. In particular, genetic drift is stronger in smaller populations, such that organisms with larger population sizes should be more diverse under neutral evolution. However, it has long been known that the amount of genetic variation does not always scale as expected with population size, with a deficit of genetic variability in larger populations as compared to the neutral expectation (Lewontin 1974). This has become known as "Lewontin's paradox." It is likely that this paradox can be explained by considering the influence of natural selection (Ohta and Gillespie 1996; Leffler *et al.* 2012; Cutter and Payseur 2013; Corbett-Detig *et al.* 2015). Since drift can act to retard selection, natural selection tends to be more efficient in organisms with larger population sizes. Consistent with this, estimated rates of adaptive evolution are often greater for smaller organisms with larger population sizes. For example, it has been estimated that >50% of amino acid substitutions between fruit fly species are driven by positive selection (Sella *et al.* 2009; Messer and Petrov 2013), but in humans, <15% of recent amino acid substitutions appear to have been driven by selection (Eyre-Walker 2006; Messer and Petrov 2013). Considering the relative importance of natural selection and genetic drift in maintaining genetic diversity has important implications for explaining current patterns of biodiversity

doi: 10.1534/genetics.115.183285
Manuscript received September 30, 2015; accepted for publication March 21, 2016; published Early Online March 24, 2016.
Available freely online through the author-supported open access option.

Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.183285/-/DC1.
[1]Corresponding author: Department of Zoology, Room S22, ,University of Cambridge, Downing St., Cambridge, CB2 3EJ, UK. E-mail: shm45@cam.ac.uk

and predicting future adaptive potential (Gillespie 2001; Leffler *et al.* 2012).

The solution to Lewontin's paradox appears to lie in the influence of natural selection on linked sites. Selection acting on one locus can cause the removal of genetic variation at physically linked, neutral loci. This can occur either through fixation of beneficial alleles ("hitchhiking") (Maynard Smith and Haigh 1974) or by purging of deleterious alleles ("background selection") (Charlesworth *et al.* 1993). Both of these processes have more pronounced effects in genomic regions of lower recombination rate. The importance of linked selection is supported by a positive correlation between recombination rate and neutral genetic diversity, first and most thoroughly studied in *Drosophila melanogaster* (Begun and Aquadro 1992; Langley *et al.* 2012; Mackay *et al.* 2012; McGaugh *et al.* 2012; Campos *et al.* 2014), and subsequently observed in other taxa, including humans (Nachman *et al.* 1998; Payseur and Nachman 2002; McVicker *et al.* 2009; Lohmueller *et al.* 2011), yeast (Cutter and Moses 2011), mice, rabbits (Nachman and Payseur 2012), and chickens (Mugal *et al.* 2013). A major recent advance has come from a population genomic analysis of 40 species, which showed not only that this phenomenon is widespread in plants and animals, but importantly, that the effectiveness of selection at removing variation at linked sites is correlated with population size (Corbett-Detig *et al.* 2015). The increased effectiveness of natural selection in larger populations reduces genetic diversity to a much greater extent than in smaller populations, countering to some degree the reduced influence of genetic drift.

However, this correlative evidence fails to capture the complexities of how natural selection acts in different species. A range of factors will affect the efficiency of natural selection and how strongly it influences linked sites, including the recombinational landscape across the genome, the frequency of adaptive change, and historical population demography (Cutter and Payseur 2013). These factors vary enormously between species, and in-depth analyses of an increasing number of taxa have revealed that not all conform to the same general trends. For example, certain plant species do not show a correlation between recombination and neutral polymorphism (see Cutter and Payseur 2013 for a thorough review). In the insects, most of what we have learned about the action of selection and drift in natural populations comes from studies of the genus *Drosophila* (Andolfatto 2007; Sella *et al.* 2009; Sattath *et al.* 2011; McGaugh *et al.* 2012; Campos *et al.* 2014; Comeron 2014; Lee *et al.* 2014). For example, in *Drosophila simulans*, genetic diversity is strongly reduced in the vicinity of recent nonsynonymous substitutions, indicative of strong hitchhiking around beneficial mutations (Sattath *et al.* 2011; Lee *et al.* 2014). This contrasts with a more subtle pattern in humans, where a reduction in diversity around functional substitutions is only detectable after accounting for background selection (Hernandez *et al.* 2011; Enard *et al.* 2014). It remains to be seen whether the rampant selection seen in *Drosophila* spp. is typical of insects.

Here we investigate the action of selection and other evolutionary forces in *Heliconius* butterflies, focusing in particular on *H. melpomene*. This species differs from *D. melanogaster* in a number of ways that might influence patterns of selection across the genome. Populations of *Heliconius* live in tropical rainforests and are characterized by long life spans and stable populations (Ehrlich and Gilbert 1973). In addition, *H. melpomene* has a similar per base recombination rate to *D. melanogaster*, but more chromosomes (21 compared to 4), potentially allowing higher overall recombination rates. Although the ecology and evolution of this genus has been the subject of much research (reviewed by Merrill *et al.* 2015), including recent genomic studies of adaptation and speciation (Arias *et al.* 2012; Nadeau *et al.* 2012, 2013; Kronforst *et al.* 2013; Martin *et al.* 2013; Supple *et al.* 2013), whole-genome studies of selection and within-species genetic diversity have been lacking. Data from *H. melpomene* was included in the recent comparative study of Corbett-Detig *et al.* (2015). However, only four individuals from a single population were considered. Here we examine in detail the action and influence of natural selection within and between populations and species using whole genome resequencing data from 59 *H. melpomene* individuals and an additional 21 samples from 11 related species. We first identify four large but cohesive populations and then explore genetic variation within and between populations and species, describing the footprints of various selective and neutral processes.

Materials and Methods

Mapping, genotyping, and estimation of error rates

The analyzed genome sequences from 80 butterflies included both published and new data. Sample information and accession numbers are given in Supplemental Material, Table S1. The 58 wild-caught *H. melpomene* samples cover much of the species range and included 13 wing pattern races. We also reanalyzed sequence data from a single individual from the inbred *H. melpomene* reference strain (Heliconius Genome Consortium 2012). For sequences generated in this study, methods were as described by Martin *et al.* (2013). All sequences analyzed here consisted of paired-end reads obtained by shotgun sequencing using either Illumina's Genome Analyzer IIx system or Illumina's HiSeq 2000 system, according to the manufacturer's protocol (Illumina).

Quality-filtered, paired-end sequence reads were mapped to the *H. melpomene* genome scaffolds (version 1.1) (Heliconius Genome Consortium 2012) using Stampy version 1.0 (Lunter and Goodson 2011). Genotypes were called using the GATK version 2.7 UnifiedGenotyper (DePristo *et al.* 2011). See File S1 for detailed methods. Only "high quality" genotype calls (Phred-scaled mapping quality and genotype quality $\geq$30) were used in downstream analyses. We optimized our genotype calling procedure by examining the total numbers of genotype calls and estimated error rates produced by different pipelines. The rate of false positive heterozygous genotype calls was estimated by analyzing a homozygous region in the inbred reference sample. We further examined how

error rates change with increasing divergence and different read depths using simulated sequence reads generated using seq-gen (Rambaut and Grass 1997) and ART (Huang *et al.* 2012). See File S1 for further details.

Analysis of phylogeny and population structure

A maximum-likelihood tree for all 80 samples was generated using only fourfold degenerate (4D) sites that had high-quality genotype calls in at least 60 samples, giving an alignment of 1.7 million bases. RAxML (Stamatakis 2006; Ott *et al.* 2007; Stamatakis *et al.* 2008) was used with the GTRGAMMA model, and 100 bootstrap replicates were performed.

We then used two approaches to identify populations that would be considered separately in downstream analyses of diversity: STRUCTURE (Pritchard *et al.* 2000; Falush *et al.* 2003), a model-based clustering method that infers the proportion of each individual's genotype made up by each of a defined number of clusters; and principle components analysis (PCA), performed using Eigenstrat SmartPCA (Price *et al.* 2006). To minimize the influence of selection, both analyses considered only fourfold degenerate sites. Detailed methods are provided in File S1.

Site frequency spectra

We generated unfolded site frequency spectra for each *H. melpomene* population by counting the number of derived alleles at biallelic sites. Sites were polarized by comparison with the "silvaniform" clade species: *H. hecale*, *H. ethilla*, and *H. pardalinus*. To allow comparison among populations, and account for missing data, each site was randomly down-sampled to the same number of individuals. See File S1 for details.

Inference of historical population size change using pairwise sequentially Markovian coalescent program

To infer changes in ancestral population sizes, we used the pairwise sequentially Markovian coalescent (PSMC) program (Li and Durbin 2011). This method fits a model of fluctuating population size by estimating the distribution of times to most recent common ancestor across a diploid genome. Twelve samples were selected *a priori* for PSMC analysis. These 12 were chosen because they all had similar sequencing depth, similar numbers of genotyped sites, were all male (homogametic, ZZ), and provided a good representation across the species range. Detailed methods are provided in File S1.

Window-based population parameters

Various population parameters were calculated for nonoverlapping 100-kb windows across the genome. Only windows with a sufficient number of sites genotyped in at least 50% of samples were considered. See File S1 for details. We used 100-kb windows because linkage disequilibrium (LD) tends to break down almost completely within 10 kb and reaches background levels within 100 kb (Figure S1), meaning that measures from adjacent windows would be largely free of linkage effects.

Nucleotide diversity (π) and absolute divergence (d_{XY}) were calculated as the average proportion of differences between all pairs of sequences, either within a sample (π) or between two samples (d_{XY}). Sites with missing data were excluded in a pairwise manner to maximize the amount of data being considered. Tajima's D (Tajima 1989) and F_{ST} (as in equation 9 of Hudson *et al.* 1992) were calculated using the EggLib Python module (De Mita and Siol 2012).

Estimating the rate of adaptive substitution

We estimated the genome-wide rate of adaptive substitution (α) using Messer and Petrov's asymptotic method (Messer and Petrov 2013), comparing synonymous and nonsynonymous SNPs covering 11,804 polymorphic genes (11,638 autosomal and 166 Z-linked). Polymorphism was measured in the Western population of *H. melpomene*, and divergence was measured between the Western population and *H. erato*. We calculated confidence intervals around the estimated α by performing 1000 bootstraps, in each of which 11,804 genes were resampled, with replacement. Detailed methods are described in File S1.

Multiple regression

We used multiple linear regression to model nucleotide diversity at 4D sites (π_{4D}) in 100-kb windows (calculated for each population separately and then averaged). The aim was to assess the influence of selection at linked sites on diversity at neutral sites. Since linked selection is largely modulated by the number of selected sites and the extent of linkage, we included as explanatory variables the local gene density (the proportion of coding sequence per window) as proxy for the density of nearby selected sites (Corbett-Detig *et al.* 2015) and local recombination rate ($\hat{r}$), calculated from the linkage map. To account for genetic hitchhiking, we also included the number of recent nonsynonymous substitutions (D_n) in the *H. melpomene* lineage per window as an explanatory variable. As an alternative, and a potentially more direct indicator of adaptive substitutions, we also tested a model using summed gene-by-gene estimates of the number of adaptive nonsynonymous substitutions (a), estimated by maximum likelihood using the McDonald–Kreitman test framework (Welch 2006). To account for mutation rate variation, the rate of synonymous substitutions per synonymous site (d_S) was also included as an explanatory variable. Lastly, we also included GC content at third codon positions to account for any effects of DNA composition. Detailed methods for the estimation of various explanatory variables and data processing for this analysis are provided in File S1.

To further investigate the interrelationships between the explanatory variables, we used principal component regression (PCR) (Drummond *et al.* 2006; Mugal *et al.* 2013). This approach can help to tease apart the effects of the various explanatory variables by summarizing the explanatory variables into orthogonal components, thereby accounting for multicollinearity. Regression analyses were performed with the R version 3.0.3 (https://www.R-project.org) using the *pls* package (Mevik and Wehrens 2007).

To assess the robustness of our findings, we investigated the influence of various modifications to the model, such as restricting the analysis to genes showing minimal codon usage bias, the exclusion of chromosome ends, and use of a different outgroup. Details are provided in File S1.

Multiple linear regression was also performed for whole chromosomes, where the response variable was the mean 4D site diversity per chromosome ($\bar{\pi}_{4D}$). Here, rather than using recombination rate estimated from the linkage map, we used chromosome length as a proxy for recombination rate (Kaback *et al.* 1992; Lander *et al.* 2001). Thus, the five explanatory variables were as follows: chromosome length, average gene density, average synonymous substitution rate ($\bar{d}_S$), average number of nonsynonymous substitutions per 100 kb ($\bar{D}_n$), and average GC content. As above, $\bar{\pi}_{4D}$ was square-root transformed, but in this model, none of the explanatory variables required transformation to correct for skewness. As above, all explanatory variables were Z-transformed so that their effects could be compared.

Scanning for selective sweeps

To identify candidate selective sweep locations in the Eastern and Western populations, we used SweeD (Pavlidis *et al.* 2013). This program is based on Sweepfinder (Nielsen *et al.* 2005) and uses a composite likelihood ratio (CLR) to identify loci showing a strong deviation in the site frequency spectrum toward rare variants. See File S1 for detailed methods.

Data availability

All raw sequence reads are available from the Sequence Read Archive from the National Center for Biotechnology Information. Accession numbers are provided in Table S1. Processed genotype data, along with data files underlying all results, figures and tables, and code used for model fitting are available from Data Dryad (http://dx.doi.org/10.5061/dryad.g0874). The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article.

Results

Genotyping

The median depth of coverage across all samples was 28×. A median of 78% of sites genome-wide, and 96% of coding sites, had high-quality genotype calls for samples of *H. melpomene*, its two close relatives *H. cydno* and *H. timareta*, and the "silvaniform" clade species *H. hecale*, *H. ethilla*, and *H. pardalinus*, which diverged from *H. melpomene* ~3.8 million years ago (MYA) (Kozak *et al.* 2015) (Figure S4). A few samples had considerably fewer sites genotyped, owing to poor sequence coverage (Table S1). More distant species, including *H. wallacei* (~8.8 MYA), *H. doris* (~9.7 MYA), and *H. erato* (~10.5 MYA) all showed strongly reduced numbers of genotype calls (median 33%), suggesting that many reads from these species were too divergent to be mapped reliably to the *H. melpomene* reference. However, the number of calls obtained in coding regions showed very little drop-off with phylogenetic distance, with a median of 90% of coding sites genotyped in the most divergent outgroup, *H. erato* (Figure S4). This implies that coding regions are sufficiently conserved to allow read mapping and genotyping across all *Heliconius* species. Our analyses of the more distant species therefore focused only on coding regions.

We selected a genotyping pipeline that gave a false positive SNP rate of 0.03% per site (three errors in 10,000 calls), when comparing the inbred reference sample to itself (Table S2). Using simulated reads, we found that our pipeline produced higher error rates for more divergent taxa, especially when sequencing depths were low (Figure S5). Nevertheless, for divergences below 6%, which is typical for coding sequences in this genus, and with appreciable sequencing depth, estimated error rates were still well under 0.05% (five in 10,000). As we are concerned primarily with large-scale genomic trends, with all analyses considering large numbers of sites, rare genotyping errors are unlikely to influence our conclusions.

Population structure and phylogenetic relationships

In order to focus our analyses on biologically meaningful populations, we first investigated population structure among our samples. Analysis of population structure based on 4D sites, using both PCA and STRUCTURE (Pritchard *et al.* 2000; Falush *et al.* 2003) gave largely congruent results. Both analyses identified three distinct *H. melpomene* clusters that were largely partitioned geographically (Figure 1, B and C). Consistent with previous studies using smaller datasets, *H. melpomene* samples from the eastern and western slopes of the Andes formed two strongly differentiated populations, separated by a deep phylogenetic split (Figure 1B). The third population was made up of the samples from French Guiana. These three populations will be referred to as the "Eastern," "Western," and "Guianan" populations. The only exception to this geographic clustering was a group of five samples of *H. m. melpomene* from the eastern slopes of the Andes in Colombia, which formed a monophyletic clade most closely allied with the Western population. However, the STRUCTURE results suggested admixture between these and both the Eastern and Guianan populations (Figure 1B). Although not differentiated by principal components 1 and 2 (Figure 1C), these five Colombian samples were differentiated from the Western population by principal component 3 (Figure S9). Given the distinct geography and genomic composition of these samples, we made the conservative decision to consider this group as a fourth distinct population ("Colombian").

Recent expansion of the Eastern population

Tests for selection can be confounded by historical changes in population size (Eyre-Walker and Keightley 2009; Li *et al.* 2012), so we investigated the population history of the four *H. melpomene* populations. Three of the four populations showed signatures consistent with fairly stable population

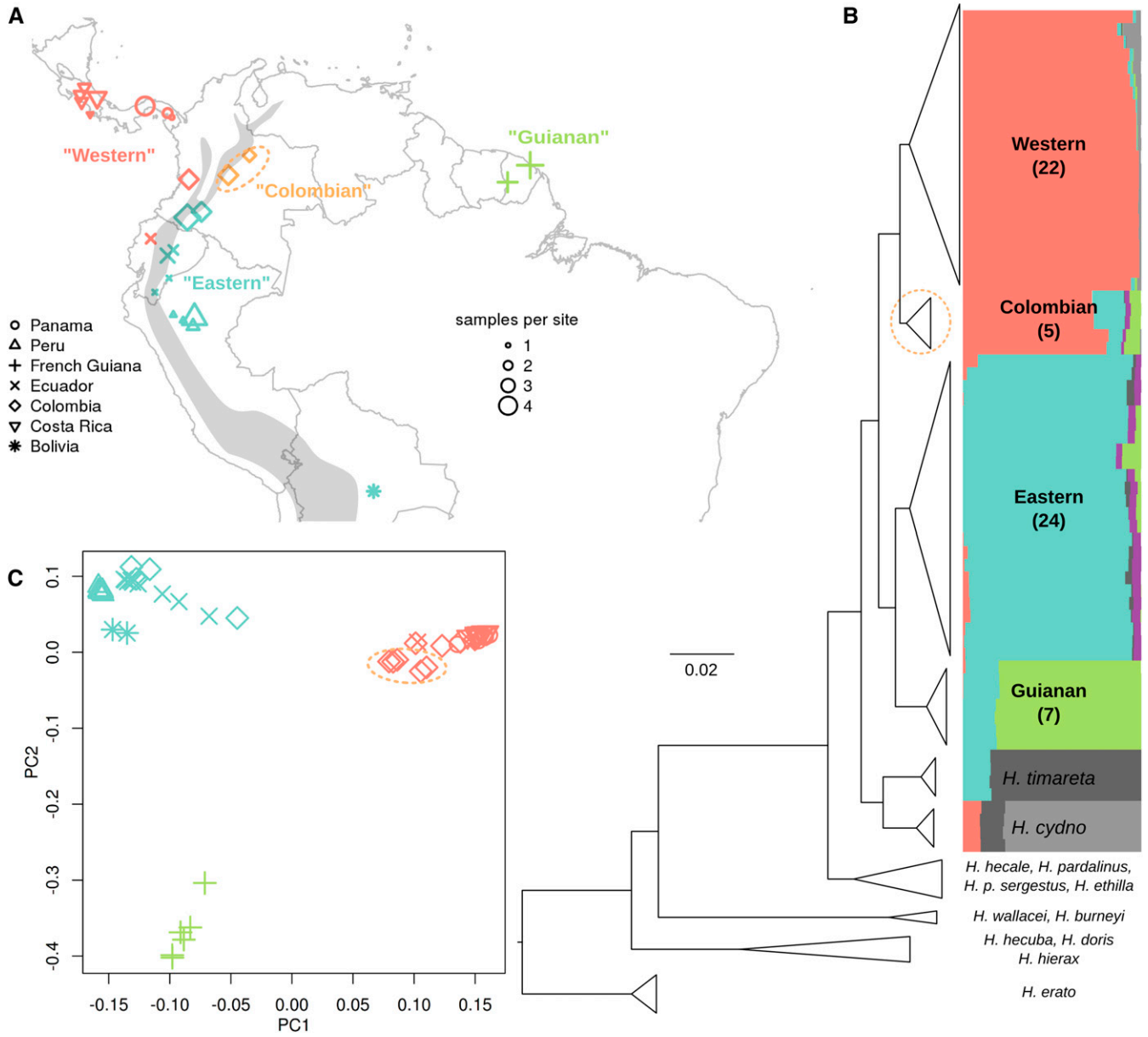

Figure 1 Sample locations, phylogeny, and population structure. (A) Sampling locations of the 58 wild *H. melpomene* samples (see Table S1 for coordinates). Symbols indicate country of sampling; sizes indicate the number of samples from each location. Colors correspond to major clustering on the STRUCTURE plot (B). Gray shading indicates the approximate location of the Andes Mountains. (B) Compressed RaxML phylogeny based on fourfold degenerate (4D) sites. See Figure S6 for an uncompressed version. Colored bars indicate genotype cluster proportions for each sample inferred by STRUCTURE with $k = 6$. STRUCTURE plots for $k = 5$–8 are given in Figure S7, and Ln probabilities for different k values are given in Figure S8. (C) Principal component 1 plotted against principal component 2, which explained 17 and 7.7% of the variance, respectively. Colors and symbols are as in A. In A–C, the Colombian samples discussed in the text are circled in orange.

sizes, but the Eastern population showed evidence of a recent expansion. Tajima's *D* was consistently negative in the Eastern population, but close to zero in the Western, Colombian, and Guianan populations (Figure 2A). Negative Tajima's *D* is indicative of an excess of rare variants, consistent with recent population growth. This finding was substantiated by direct examination of the unfolded site frequency spectrum (SFS). The SFS at 4D sites showed a strong excess of rare variants in the Eastern population compared to the other populations (Figure 2B). This skew remained when only geographically proximate samples, with high sequencing coverage, were considered (Figure S10), indicating that it was not an artifact of sampling design. The same trend was also observed at intronic and intergenic sites (Figure S10). Compared to neutral expectations with constant population size, the other three populations displayed a weak excess of rare variants, but to a much lesser degree than the Eastern population. For the Eastern and Western populations, which were more densely sampled, we were able to compare the SFS down sampling to 20 individuals for each SNP position. This deeper

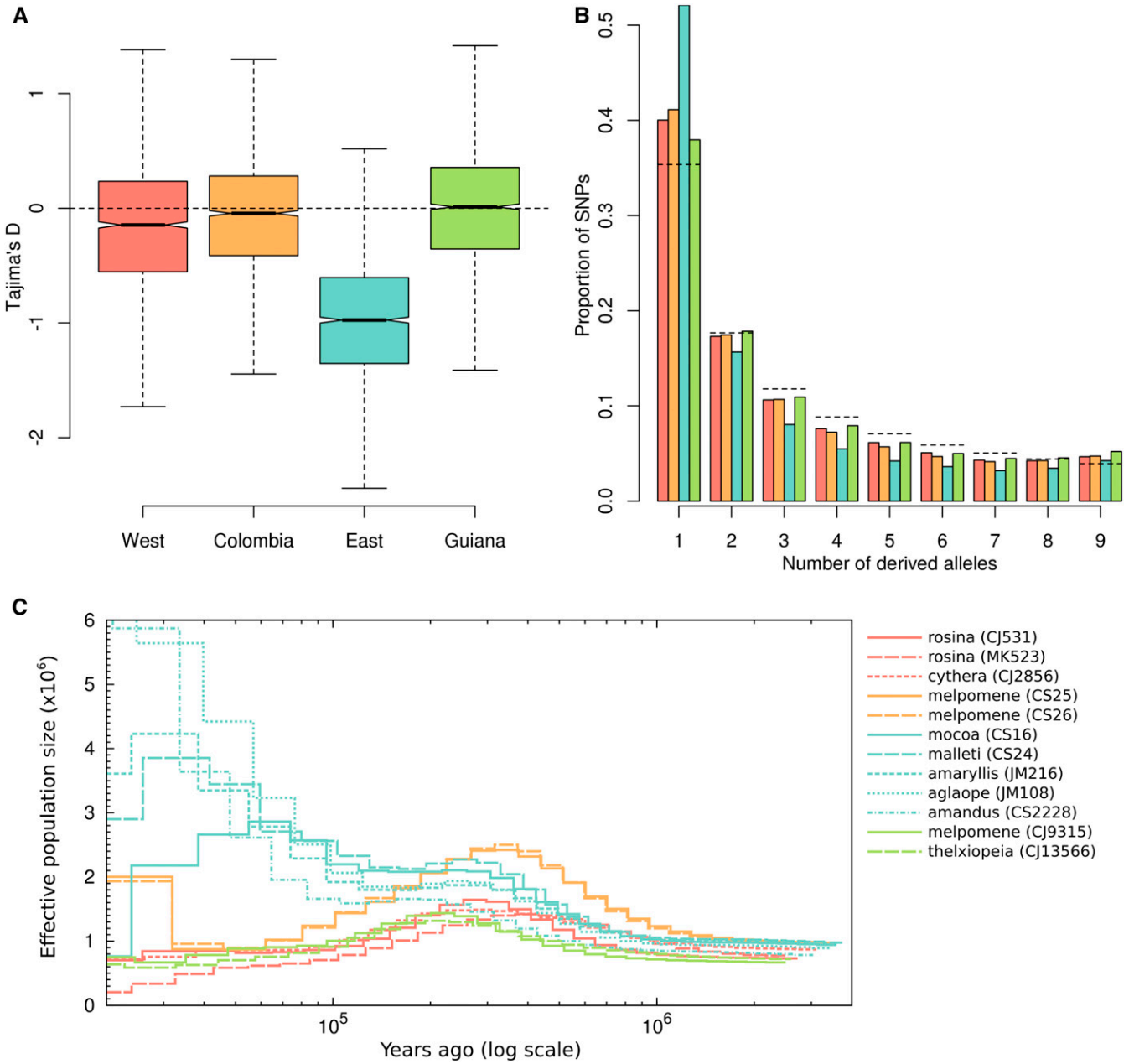

Figure 2 Evidence for recent expansion of the Eastern population. (A) Boxplots of Tajima's *D*, calculated for 4D sites in each 100-kb window throughout the genome. (B) Site frequency spectra for fourfold degenerate (4D) sites for the four *H. melpomene* populations, sampling five individuals per site. Colors denote populations: red, Western; orange, Colombian; blue, Eastern; and green, Guianan. Dashed lines indicate the expected frequencies under the standard coalescent model with constant population size (Fu 1995). (C) PSMC plots of inferred population size through time on a logarithmic axis. Twelve selected male samples that had similar numbers of genotyped sites were included. Source populations are colored as in A and B.

sampling reproduced the pattern, further showing that the skew was not limited to singleton variants, but also doubletons (derived alleles present twice in the sample) (Figure S10). While genotyping error could explain some of the excess of singleton SNPs, it is unlikely to cause the observed excess of doubletons, nor the dramatic skew seen in the Eastern population.

We further verified our hypothesis of a recent expansion in the Eastern population using the PSMC method of Li and Durbin (2011). All four populations showed similar population size histories up until ~200,000 years ago, with a gradual increase in population size beginning ~1 MYA and leveling off ~300,000 years ago. However, while the Western, Colombian, and Guianan samples showed a subsequent decrease in the inferred N_e, that of the Eastern samples rose again, roughly doubling between 100,000 and 30,000 years ago (Figure 2C). Closer to the present (<30,000 years ago) the inferred individual histories diverged considerably, as may be expected given the dearth of information about recent demography to be gained from analysis of single

Table 1 Nucleotide diversity (π) in *H. melpomene* and absolute divergence (d_{XY}) from outgroups

		d_{XY} between *melpomene* and:				
Site class	π (*melpomene*)	*cydno* and *timareta*	*hecale*, *ethilla*, and *pardalinus*	Wallacei and *burneyi*	*doris*, *hecuba*, and *hierax*	*erato*
Whole genome						
All sites	0.019	0.027	0.036	.	.	.
Intergenic	0.020	0.029	0.038	.	.	.
Intron	0.019	0.028	0.038	.	.	.
Codon 1	0.006	0.010	0.014	0.026	0.022	0.037
Codon 2	0.006	0.009	0.012	0.022	0.020	0.032
Codon 3	0.015	0.024	0.033	0.065	0.056	0.091
4D	0.025	0.041	0.057	0.114	0.100	0.158
Z chromosome						
All sites	0.011	0.024	0.034	.	.	.
Intergenic	0.012	0.025	0.035	.	.	.
Intron	0.011	0.025	0.036	.	.	.
Codon 1	0.003	0.008	0.013	0.028	0.024	0.036
Codon 2	0.003	0.007	0.012	0.023	0.021	0.030
Codon 3	0.007	0.019	0.030	0.064	0.059	0.094
4D	0.011	0.032	0.049	0.107	0.100	0.158

See Table S3 for full data including error margins.

genomes (Li and Durbin 2011). Nevertheless, there appears to be a tendency for the more southern samples to show greater expansion (*e.g.*, *H. m. amandus* from Bolivia and *H. m. aglaope* from Peru). Methods more sensitive to demographic change in the recent past may be necessary to confirm this pattern. This analysis was repeated several times, varying the PSMC input parameters for block size and recombination rate. While absolute population size estimates tended to be lower at smaller block sizes, the observed trend of a population size expansion in each of the Eastern samples was consistent throughout (data not shown). As it is based on heterozygosity in single genomes, this analysis is independent of the site frequency spectrum and therefore provides an additional line of evidence for a recent expansion of the *H. melpomene* population to the east of the Andes.

One potential caveat in this conclusion is that hybridization and gene flow may also produce patterns consistent with population growth, and gene flow is known to occur between the eastern population and *H. timareta* (Martin *et al.* 2013). However, there is also significant gene flow between the Western population and *H. cydno*, implying that hybridization alone is unlikely to explain the distinct pattern seen in the Eastern population.

Diversity and divergence across the genome

Estimated neutral diversity in *H. melpomene* was found to be high, and comparable with that in *Drosophila* spp. Estimates of within-population nucleotide diversity (π) in *H. melpomene* made use of only those samples with at least 25× depth of coverage, because we found that levels of within-sample heterozygosity tended to be underestimated at sequencing depths below this threshold (Figure S11). Genome-wide π, averaged over all 100-kb windows across the four populations was 1.9%, and similar when only intergenic (2.0%) or intronic (1.9%) sites were considered (Table 1; Table S3). As expected, diversity was strongly reduced at first and second codon positions (0.6%) and higher at third codon positions (1.5%). Diversity was highest at 4D sites (2.5%).

The peripheral Western and Guianan populations had significantly lower diversity than those at the center of the range (Figure 3A) (paired Wilcoxon signed rank test, $P < 2e\text{-}16$). To ensure that this trend was not simply driven by population substructure among sampled individuals, we examined levels of heterozygosity within each sample. As mentioned above, this revealed that sequencing depth affected estimates of heterozygosity, but that above a depth of roughly 25×, heterozygosity was consistent within each population. Considering only samples with depth of at least 25×, we found that average 4D site heterozygosity in the Western samples (2.67%) was only marginally lower than that in the Colombian (2.79%) and Eastern (2.82%) samples, whereas that of the Guianan samples (2.16%) remained considerably lower than the other populations (Figure S11).

To estimate as closely as possible the value of $\theta = 4N_e\mu$, we recalculated average diversity at 4D sites considering only autosomal genes showing minimal codon usage bias. This gave a slightly higher value of 2.7%, ranging from 2.1 to 2.9% across the four populations (Table 2). These values are in the same range as estimated neutral diversity for *D. melanogaster* in Southern Africa (~2%) and *D. simulans* (~3.5%) (Begun *et al.* 2007; Langley *et al.* 2012).

Mean divergence at 4D sites between *H. melpomene* and *H. erato*, its most distant relative in the genus, was 15.8% [16% when only minimal codon usage bias (CUB) genes were considered]. A calibrated phylogeny places the split between these two species at ~10.5 MYA (Kozak *et al.* 2015). Assuming four generations per year, this corresponds to a neutral mutation rate of 1.9×10^{-9} per site per generation. This is about two-thirds of the spontaneous mutation rate recently estimated using whole genome sequencing of parents and

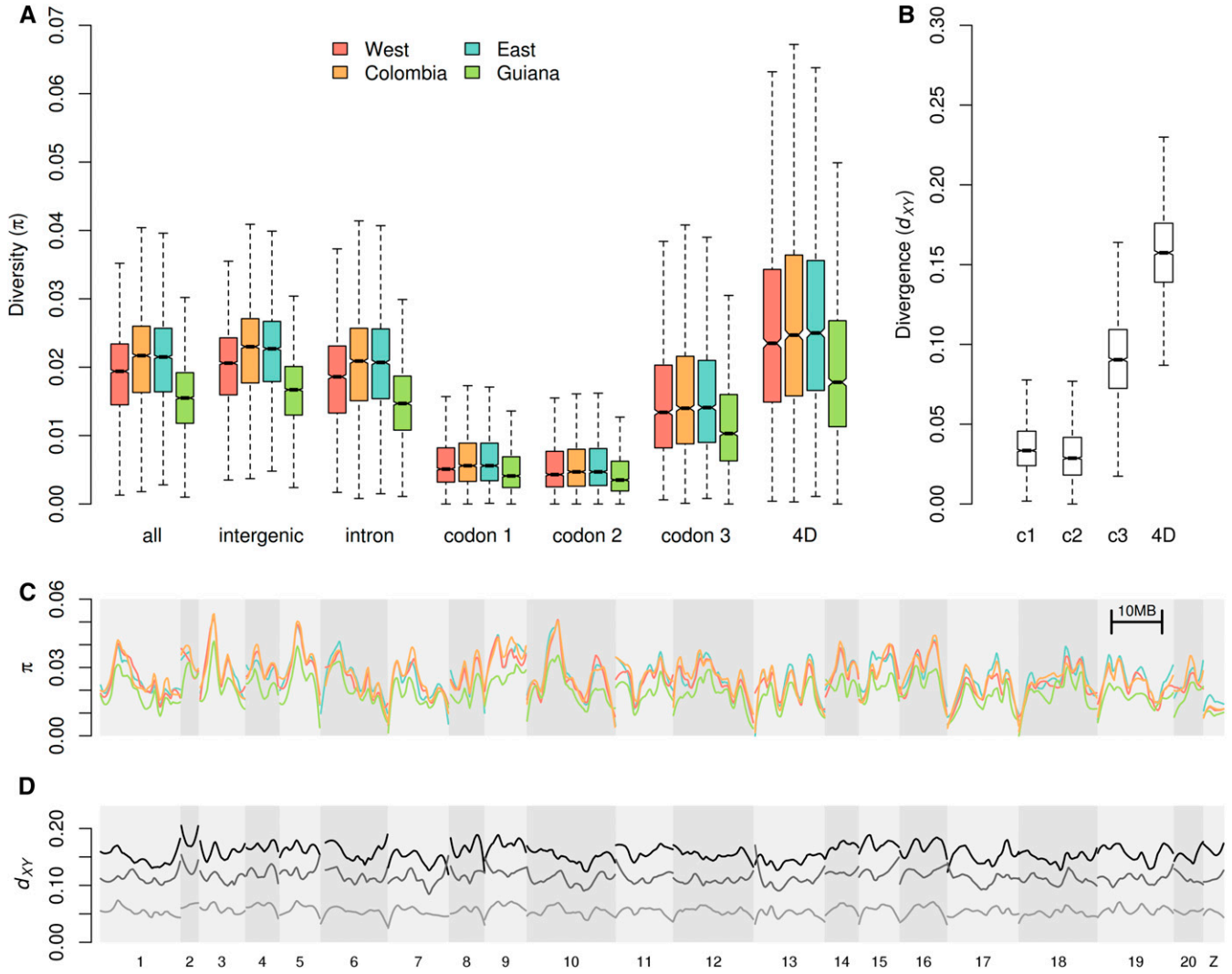

Figure 3 Genome-wide diversity and divergence. (A) Boxplots of nucleotide diversity (π) for different site classes in the four *H. melpomene* populations. π values were calculated in 100-kb windows, considering all sites of each class within each window. (B) Boxplots of divergence (d_{XY}) between *H. melpomene* and *H. erato* at four site classes: first, second, and third codon positions and fourfold degenerate (4D) sites. Note the different *y*-axis scale. (C) Nucleotide diversity at 4D sites plotted across the 21 *H. melpomene* chromosomes (shaded). Scaffold order was inferred from the *H. melpomene* genome linkage map of v1.1. Populations are colored as in A. Values are for nonoverlapping 100-kb windows, smoothed with loess (local regression), with a span equivalent to 4 Mb. (D) Divergence (d_{XY}) across the genome, between *H. melpomene* and the silvaniform clade, *H. doris* clade, and *H. erato* (respectively, from light to dark), smoothed as in C.

offspring in *H. melpomene* (2.9×10^{-9}) (Keightley *et al.* 2014). Using these two rates, we estimated N_e ($\theta/4\mu$) for the four populations, which ranged from 1.8–2.8 million for the Guianan population to 2.5–3.8 million for the Colombian population (Table 2). We note that these do not represent instantaneous values, but rather aggregates over the course of the coalescent time scale. They are also based on a small subset of putatively neutral sites, making them difficult to compare with the PSMC results (Figure 2C), as the latter are based on the whole genome data, for which the level of polymorphism is ~30% lower (Table 1).

Levels of diversity at 4D sites varied considerably across individual chromosomes (Figure 3C). This heterogeneity was strongly conserved between the three populations. Interspecific divergence also varied across the chromosomes, but to a lesser extent (Figure 3D). Diversity was also reduced on the Z chromosome relative to autosomes, as expected, given its lower effective population size (N_e). This discrepancy between autosomes and Z was also present in measures of divergence between *H. melpomene* and its closer relatives, but disappeared at higher levels of divergence, with d_{XY} between *H. melpomene* and *H. erato* being nearly identical for autosomes and Z (Table 1). This is consistent with a decreasing contribution of N_e to coalescence time for deeper species splits.

One potential concern is that highly variable regions may have been missed due to poor read mapping, in which case diversity and divergence might be underestimated. To test for

Table 2 Estimated neutral π (θ) for the four populations, and corresponding population size estimates (in millions) given two different mutation rates

Population	Neutral π (θ)	N_e (×10⁶) [μ = 2.90 × 10⁻⁹]	N_e (×10⁶) [μ = 1.90 × 10⁻⁹]
Western	0.028	2.385	3.641
Colombia	0.029	2.503	3.821
Eastern	0.029	2.469	3.769
Guiana	0.021	1.820	2.78

this bias, we investigated whether π and d_{XY} were correlated with the proportion of missing data per window (measured as sites genotyped in fewer than 50% of samples). Third codon positions averaged just 2.3% missing data among the *H. melpomene* samples, and just 2.4% across the entire set of 79 wild samples. Neither π nor d_{XY} were correlated with the proportion of missing data (Figure S12, Figure S13). Hence, there is no evidence for any bias in coding regions. In intergenic regions, which averaged 23% missing data in *H. melpomene* samples and 25% across the whole sample set, both π and d_{XY} were found to be weakly correlated with the proportion of missing data (Figure S12). However, the amount of variance explained by missing data was very low (linear regression, $R^2 = 0.037$ and 0.009, respectively). The effect of missing data on our estimates of diversity and divergence therefore appears to be minimal. We suggest that the reduced rate of read mapping in noncoding regions in *H. melpomene* and its closer relatives is probably driven more by an abundance of repeats and structural variation rather than by excessively divergent sequences.

The rate of adaptive fixation

We estimated that 31% of fixed amino acid substitutions between *Heliconius* species are adaptive. We used Messer and Petrov's asymptotic method (Messer and Petrov 2013) to estimate a genome wide α, the proportion of nonsynonymous substitutions driven by positive selection. The exponential model showed a good fit to the data (Figure 4), and gave an estimated α of 31.0%. The 5th and 95th quantiles from 1000 bootstrap replicates were 29.0 and 33.0%, respectively. This value is roughly intermediate between estimates for humans (13%) and *D. melanogaster* (57%), generated using the same approach.

Selection reduces diversity at linked neutral sites

We explored the influence of selection on linked sites using a multiple linear regression approach, following several previous studies (Cutter and Moses 2011; McGaugh *et al.* 2012; Mugal *et al.* 2013). Our "main model" had nucleotide diversity at 4D sites (π_{4D}) in 100-kb windows as the response variable and five explanatory variables: (i) local gene density, a proxy for the number of nearby selected sites; (ii) local recombination rate ($\hat{r}$); (iii) the number of recent nonsynonymous substitutions (D_n), to account for hitchhiking around beneficial mutations; (iv) the synonymous substitution rate

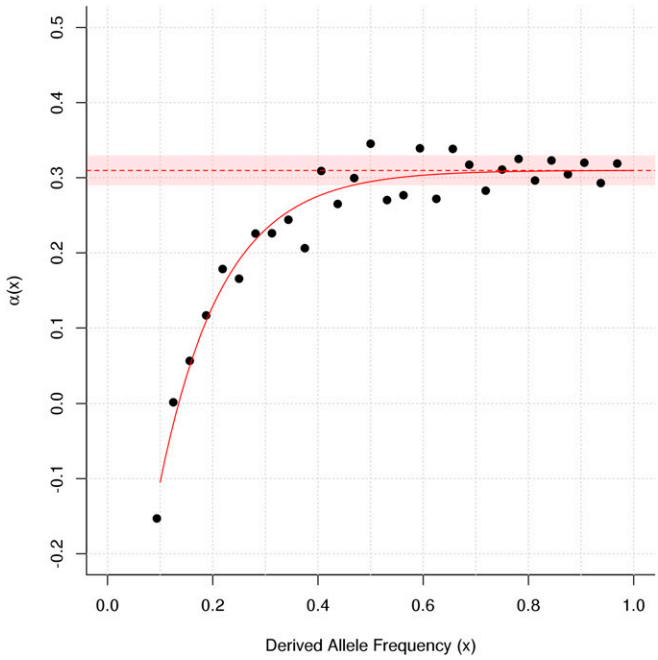

Figure 4 Estimating the rate of adaptive substitution. Mean genome-wide α was estimated using the "asymptotic" method (Messer and Petrov 2013), based on polymorphism for the Western population of *H. melpomene*, and divergence between the Western population and *H. erato*. α(*x*) was calculated for each derived allele frequency (*x*) for all $x \geq 0.1$. The solid red line indicates the fit of the asymptotic exponential function $\alpha(x) = a + b\exp(-cx)$, extrapolated to $x = 1$ (dashed red line). The solid red box indicates the range between the 5th and 95th percentiles from 1000 bootstrap samples over the 11,804 genes used.

(d_S), a proxy for local mutation rate; and (v) GC content, to account for effects of local DNA composition (Figure S14, Figure S15). The results for the main model are summarized in Table 3 and Table S4 and described below. There was limited serial correlation among windows. The Durbin–Watson statistic (Durbin and Watson 1950, 1951) for the main model was 1.3, suggesting that autocorrelation is unlikely to influence our conclusions (Field 2009). Several modifications of the model were also tested to investigate the robustness of the result. These are all summarized in Table S5, Table S6, Table S7, Table S8. Overall, results were consistent throughout, but several notable differences are described below.

The main model explained 34% of the variation in π_{4D} ($F_{5,1461} = 151.1$, $P < 2.2e\text{-}16$) and all five predictor variables were found to have significant effects (Table 3). Unsurprisingly, synonymous substitution rate (d_S) was a strong predictor of intraspecific diversity ($F_{1,1461} = 260.89$, $P < 2.2e\text{-}16$), implying that at least some of the observed heterogeneity in diversity across the genome is explained by variation in mutation rate. Local gene density also showed a strong negative relationship with π_{4D} ($F_{1,1461} = 162.03$, $P < 2.2e\text{-}16$), consistent with a considerable effect of selection at linked sites. Local recombination rate showed a positive relationship with diversity ($F_{1,1461} = 65.27$, $P < 1.357e\text{-}15$), also as expected under linked selection. In addition, recombination rate was

Table 3 Summary of multiple regression with five explanatory variables for 4D site diversity

Variable	Estimate	SE	SS	RSS	$F_{1,1461}$	P (>F)	Partial R^2	VIF
Gene density	−0.0127	0.0010	122.876	1230.8	162.028	<2.2e-16*	0.0998	1.483
$\hat{r}$	0.0062	0.0008	49.496	1157.5	65.267	1.357e-15*	0.0428	1.011
D_n	−0.0038	0.0010	11.432	1119.4	15.075	0.0001*	0.0102	1.567
d_S	0.0147	0.0010	197.848	1305.8	260.889	<2.2e-16*	0.1515	1.156
GC content	−0.0036	0.0008	14.391	1122.4	18.977	1.416e-05*	0.0128	1.076
(Intercept)	0.1516	0.0009						

Calculated for 100-kb windows (R^2 = 0.340; adjusted R^2 = 0.339; $F_{5,1461}$ = 151.1; P < 2.2e-16). $\hat{r}$, recombination rate; D_n, number of nonsynonymous substitutions; d_S, synonymous substitutions per synonymous site; SS, sum of squares; RSS, residual sum of squares; VIF, variance inflation factor; *$P \leq 0.05$.

not correlated with d_S (Pearson's $R = -0.01$, $P < 0.648$; Figure S16), implying that the positive relationship with diversity cannot be explained by a mutagenic effect of recombination. We suggest that the true relationship between recombination rate and neutral diversity may be stronger than our model predicts, as the imperfect placement of scaffolds on the Hmel1.1 linkage map limits our ability to accurately estimate local recombination rates (see below).

Evidence for genetic hitchhiking

There was also a significant negative effect of the number of nonsynonymous substitutions per window (D_n) on π_{4D} ($F = 15.08$, $P < 0.0001$), implying a small but detectable effect of genetic hitchhiking around adaptive substitutions. While multicollinearity among the explanatory variables was generally weak, there was a moderate correlation between gene density and D_n (Pearson's $R = 0.53$; Figure S16). This is unsurprising, since nonsynonymous substitutions can occur only in coding sequence and are thus likely to be more common in gene-rich regions. This can make it difficult to separate the roles of hitchhiking and background selection. However, the variance inflation factors for gene density and D_n in our model were low and do not suggest a significant impact of collinearity on the precision of our estimates (Table 3). Moreover, PCR analysis, which first accounts for collinearity among explanatory variables by separating them into principle components before performing a multiple regression, demonstrated an effect of gene density on diversity independent of D_n (Figure S17).

One potential concern is that D_n could be underestimated in highly divergent genes, where read mapping for the distant outgroup *H. erato* may be poor. We therefore tested a modified model in which d_S and D_n were estimated using only the more closely related silvaniform species as outgroups. This made little difference to the results (Table S5).

As an alternative to D_n, we also tested a modified model that included maximum-likelihood estimates of a (the number of adaptive nonsynonynmous substitutions) per gene, summed across each window. This revealed a similar significant negative effect of a on π_{4D} (Table S6), while collinearity with gene density was considerably lower than for D_n (Pearson's $R = 0.27$). Taken together, these finding all support a role for genetic hitchhiking in addition to background selection in shaping neutral variation in *H. melpomene*.

One striking difference between humans and fruit flies is that genetic hitchhiking in *Drosophila* spp. is pervasive enough to produce an average trough in diversity around nonsynonymous substitutions (after scaling for mutation rate variation) (Sattath *et al.* 2011; McGaugh *et al.* 2012); whereas this is not directly observable in humans (Hernandez *et al.* 2011). We performed an equivalent test with our data (Figure S18), and found patterns similar to those in humans, with no significant reduction of scaled diversity in the vicinity of nonsynonymous substitutions compared to synonymous substitutions. Enard *et al.* (2014) suggest that the effect of hitchhiking around nonsynonymous substitutions can be masked by background selection. This may be because background selection tends to be stronger in more conserved genomic regions, where adaptive substitutions are expected to be less common. This might explain the fact that we observe evidence for reduced diversity only around nonsynonymous substitutions in our multiple-regression model. Although our approach does not model the action of background selection explicitly, by including gene density and recombination rate as explanatory variables, it may account for some of the confounding influence of background selection. The relative roles of hitchhiking and background selection may be further resolved in the future by explicitly modeling these different processes (Corbett-Detig *et al.* 2015).

GC content and codon usage bias

In the main model, GC content was found to be negatively correlated with π_{4D} ($F_{1,1461} = 18.98$, $P = 1.416e-05$). This may reflect CUB, with a preference for codons ending in C or G, which would lead to elevated GC content at genes under stronger selection for codon usage (Wright 1990). Indeed, in our analysis of codon usage, genes with a higher GC content at the third codon position tended to display stronger evidence for CUB (Figure S5). In a modified model where π_{4D} was calculated using only our defined set of minimal CUB genes, GC content became a nonsignificant predictor of diversity, whereas effect sizes for all other explanatory variables were similar (Table S7).

Effect of chromosome ends

We last confirmed that the observed patterns are not predominantly driven by chromosome ends. A model excluding

windows in the outer 5% of chromosomes was similar to the main model (Table S8). Generally, effect sizes and *P*-values were lower, but this may partly reflect the 10% reduction in the number of observations. Interestingly, GC content was no longer a significant predictor of diversity. This might imply that patterns of codon usage change toward the chromosome ends, but this will require further investigation.

The relationship between gene density and neutral diversity is clearly visible

Given the large effect of gene density on diversity at 4D sites, we further explored this relationship visually (Figure 5). There was a clear trend of lower π_{4D} in gene-rich regions, but also a conspicuous increase in variance in regions of lower gene density (Figure 5A). This is most likely caused by the smaller number of 4D sites available in such regions, resulting in fewer data and therefore increased noise. We were able to account for this issue in our multiple linear regression model by weighting residuals according to the number of data points available per window, and overall the model showed minimal violation of assumptions (Figure S14, Figure S15). On several chromosomes, the correlation between gene density and π_{4D} was remarkably clear (Figure 5B).

Longer chromosomes are less polymorphic

There was a strong negative relationship between 4D site diversity and chromosome length (in bases) (Table 4), further supporting the pervasive role of linked selection in shaping genetic diversity in *H. melpomene*. Long chromosomes tend to have lower recombination rates per base pair (Kaback *et al.* 1992; Lander *et al.* 2001; Kawakami *et al.* 2014), which should lead to stronger linked selection. Although a considerable number of scaffolds in the *H. melpomene* v1.1 genome are not properly positioned and oriented on a chromosome, the chromosomal assignment could be inferred for almost all large scaffolds (83% of the genome in terms of bases) (Heliconius Genome Consortium 2012), making for fairly robust estimates of chromosome length. We used a multiple regression model similar to that used for 100-kb windows above, but here averaging all parameters over each of the 20 autosomes, and with chromosome length used as a proxy for recombination rate. This model explained 73.34% of the variation in average chromosomal diversity at 4D sites, $\bar{\pi}_{4D}$ ($F_{5,14} = 7.7$, $P = 0.0011$) (Table 4). There was a strong negative relationship between $\bar{\pi}_{4D}$ and chromosome length. Comparing models with and without chromosome length as an explanatory variable, we found that the model including chromosome length had a far better fit to the data ($F_{1,15} = 20.15$, $P < 0.0005$). Hence, long chromosomes tend to be less variable at neutral sites than short chromosomes, and this trend was clear upon visual inspection (Figure 6A). As in the window-based model, $\bar{\pi}_{4D}$ was positively correlated with average synonymous substitution rate, $\bar{d}_S$ ($F_{1,14} = 9.85$, $P = 0.0073$), but there was no significant relationship between $\bar{d}_S$ and chromosome length (Pearson's $r = 0.08$, $P = 0.7289$) (Figure 6B), reinforcing our finding that mutation rates are not correlated with recombination rate. The simplest explanation for this pattern is therefore that linked selection drives patterns of diversity not only among small windows, but also among whole chromosomes.

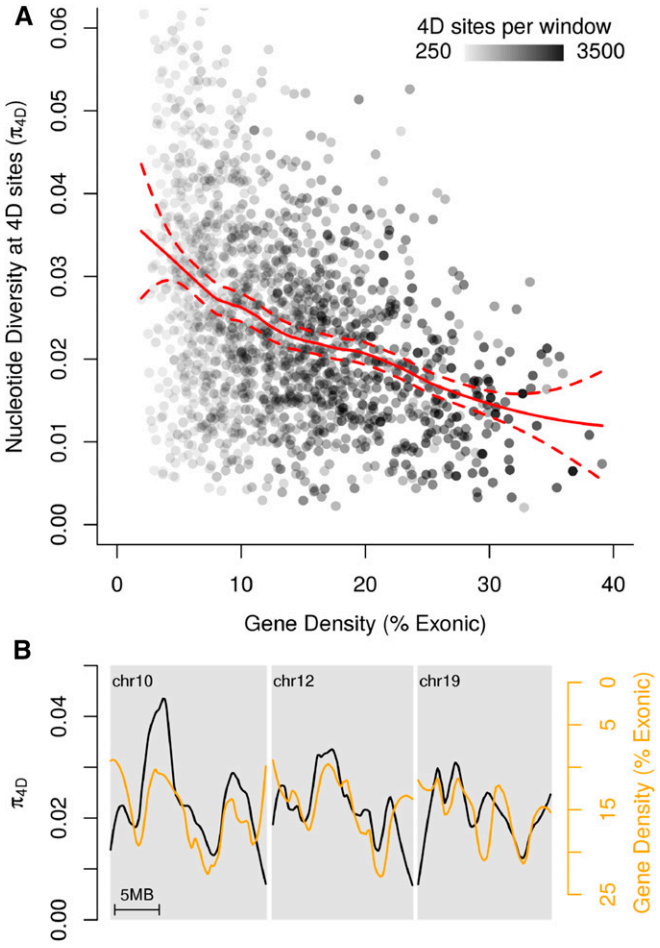

Figure 5 Relationship between gene density and diversity at 4D sites. (A) Diversity at 4D sites (π_{4D}) for 100-kb windows plotted against local gene density, calculated as the percentage of the window made up of exons. Points are shaded according to the number of 4D sites in the window that had genotype calls for at least 50% of analyzed samples. A loess (locally weighted smoothing, span = 0.5) curve with 99% confidence intervals is shown in red. (B) Plots of π_{4D} (black) and gene density (orange) across chromosomes 10, 12, and 19, which all showed a visually striking correlation. Note that the gene-density axis is inverted and adjusted to aid comparison between the lines. Both lines are loess smoothed with a span equivalent to 4 Mb.

The only other noteworthy correlation with chromosome length was a negative relationship with GC content. We hypothesized that this skew in base composition may be driven by stronger CUB on shorter chromosomes. This would be expected if higher recombination rates on shorter chromosomes allowed for more efficient selection by reducing interference among sites (Hill and Robertson 1966; Felsenstein 1974). Consistent with this hypothesis, the proportion of genes identified as having nonnegligible CUB was negatively correlated with chromosome length (Spearman's $r = -0.555$, d.f. = 19, $P = 0.009$) (Figure S19).

Table 4 Summary of multiple regression with five explanatory variables for mean 4D site diversity

Variable	Estimate	SE	SS	RSS	$F_{1,14}$	P (>F)	Partial R^2	VIF
Gene density	0.0001	0.0007	1.960e-07	8.020e-05	0.034	0.8557	0.0024	1.425
Chr length	−0.0034	0.0008	1.152e-04	1.952e-04	20.153	0.0005*	0.5901	1.920
$\bar{D}_n$	−0.0032	0.0013	3.318e-05	1.132e-04	5.807	0.0303*	0.2932	5.990
$\bar{d}_S$	0.0036	0.0012	5.628e-05	1.363e-04	9.849	0.0073*	0.4130	4.495
GC content	−0.0019	0.0011	1.815e-05	9.815e-05	3.176	0.0964	0.1849	3.722
(Intercept)	0.0257	0.0005						

Calculated for chromosomes (R^2 = 0.7334; adjusted R^2 = 0.6382; $F_{5,14}$ = 7.704; $P < 0.001144$). $\bar{D}_n$, average number of nonsynonymous substitutions per 100 kb; $\bar{d}_S$, average synonymous substitutions per synonymous site; SS, sum of squares; RSS, residual sum of squares; VIF, variance inflation factor; *$P \leq 0.05$.

Geographically restricted selective sweeps

We investigated whether there were signatures of strong, recent selective sweeps in *H. melpomene* and also whether sweeps tended to be geographically restricted to a particular population. We scanned the genome for putative selective sweep signals using SweeD (Pavlidis *et al.* 2013), which uses the CLR method of Nielsen *et al.* (2005) to identify loci displaying a strong skew in the SFS toward rare variants in comparison with the genomic background. A number of regions throughout the genome had outlying CLR values, above a threshold determined using neutral simulations (Figure S20). Most of these were restricted to the Eastern population, with only one being restricted to the Western population, and one region with partially overlapping outliers in both populations. Given the excess of outliers in the Eastern population, most of which are represented by just single windows, it seems likely that most of these signals are spurious, perhaps reflecting increased variance in the SFS caused by a less stable demographic history. Only two loci had strong putative sweeps indicated by clusters of outlying CLR values. On chromosome 11, outliers in the Eastern and Western populations roughly overlapped (Figure 7A), consistent with a beneficial allele spreading across both populations. On chromosome 12, the cluster of outlying CLR values was restricted to the Eastern population (Figure S20). We investigated patterns of diversity within and divergence between populations at these two loci in finer detail. The results were very similar for both candidate sweep loci, so we present the results for chromosome 11 in Figure 7, and those for chromosome 12 in Figure S21.

At the putative sweep locus on chromosome 11, nucleotide diversity (π) in both the Eastern and Western populations was reduced (Figure 7B). Absolute divergence between the two populations (d_{XY}) was similarly reduced (Figure 7B). However, the fixation index, F_{ST}, between the Eastern and Western populations was strongly elevated in this region (Figure 7C). This implies that the region carries a low overall amount of genetic variation, but that the variation that is present constitutes fixed or nearly fixed differences between the two populations (Charlesworth 1998). This is consistent with a scenario where the alleles that swept to high frequency in the two populations were not identical, but similar, and potentially carried the same beneficial allele (Bierne 2010). The putative sweep locus on chromosome 12 showed very much the same pattern (Figure S21).

To further explore the genetic make-up of these regions, we visualized the genotypes of individuals from all four populations at a sample of 600 highly polymorphic biallelic SNPs across the scaffold containing the putative sweep locus. Both scaffolds had regions in which heterozygous genotypes were clearly reduced, and fixed differences between the Eastern and Western populations strongly increased. In both cases, there were also several fixed differences between the Eastern and Guianan populations, but no fixed differences between the Western and Colombian populations among the 600 sampled SNPs. We note that by focusing on highly polymorphic SNPs, we highlight differentiation between the populations and fail to show how much of the region is shared, which must be considerable, given the reduced d_{XY}. Nevertheless, this visualization confirms our hypothesis that distinct alleles reached high frequency in the different populations. The presence of some heterozygous sites in the sweep regions indicates that these may be fairly ancient sweeps and/or that no single allele fixed in each population (*i.e.*, a "soft sweep"). One notable observation is the presence of long runs of heterozygous genotypes in certain individuals. This is also consistent with a soft sweep, but may alternatively reflect gene flow subsequent to the sweep, leading to long haplotypes introgressing between the populations.

Both putative sweep locations contained multiple annotated genes. All genes in these two regions that gave significant BLAST hits to *D. melanogaster* proteins (18 and 13 genes, respectively) are listed in Table S9. Due to the large number of potential targets of selection, we do not speculate here as to the adaptive significance of these events.

Discussion

The extent to which evolutionary change is a result of neutral or selective forces remains one of the long-standing questions in evolutionary biology. Genomic data permit powerful tests of the various forces that shape genetic variation within and between species, but such studies have only recently been extended beyond a few well-studied taxa. We examined a large number of whole genome sequences to investigate the forces shaping diversity in *Heliconius* butterflies. Levels of neutral diversity in *H. melpomene* are similar to those in Southern African populations of *D. melanogaster*, suggesting comparable effective population sizes. However, actual census

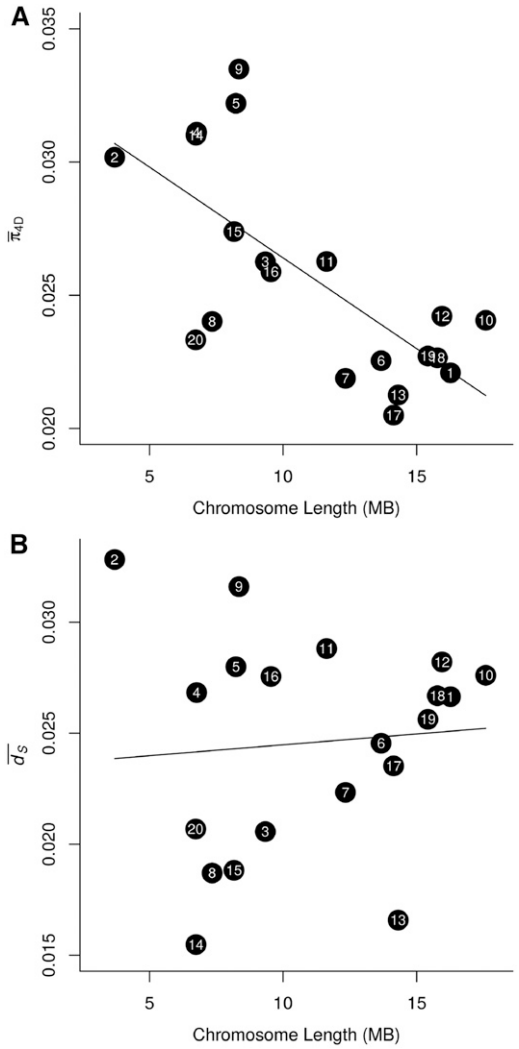

Figure 6 Relationships between chromosome size, 4D site diversity, and synonymous substitution rate. (A) Average 4D site diversity per chromosome ($\overline{\pi}_{4D}$) plotted against chromosome length. (B) The average rate of synonymous substitution per synonymous site per ($\overline{d}_S$) plotted against chromosome length. In both plots, chromosome numbers are indicated, and a linear model fit is shown for reference.

population sizes of fruit flies must at times reach numbers far greater than those of *Heliconius* butterflies, which are characterized by low-density, stable populations (Ehrlich and Gilbert 1973). This paradox of divergent demography but similar diversity is partly explained by the impact of selection on linked sites. Rampant selection across the compact *Drosophila* genome leads to a dramatic reduction in diversity at linked sites. By contrast, our results suggest that selection is less pervasive in *Heliconius*, and its influence on linked sites, though significant, is less pronounced.

Our findings suggest that positive selection plays an important, but less prominent role in *Heliconius*. We estimate that 31% of amino acid substitutions between *H. melpomene* and *H. erato* were driven by positive selection. This contrasts with an estimated rate of adaptive substitution in *D. melanogaster* of 57%, made using the same method (Messer and Petrov 2013). Although our multiple-regression model indicated that genetic hitchhiking has had a significant effect on neutral variation, we did not detect the same strong reduction in diversity around nonsynonymous substitutions seen in *Drosophila* spp. (Sattath *et al.* 2011; McGaugh *et al.* 2012). One possible explanation is that positive selection more often targets noncoding regulatory changes, as appears to be the case in humans (Enard *et al.* 2014). However, even more general signatures of recent selective sweeps in the form of strong skews in the site frequency spectrum were limited. It is important to note that the amount of adaptive evolution depends not just on the efficacy of selection, but also on how often novel, adaptive phenotypes arise. This depends both on the changeability of the fitness landscape and the availability of adaptive variation. Populations of fruit flies may occasionally reach extremely high densities, increasing the potential for selection to detect advantageous mutations (Barton 2010). The relevant population size for adaptive evolution could therefore be much higher than that estimated from neutral variation, especially in species with highly fluctuating populations.

Certain aspects of *Heliconius* biology might also obscure the footprints of positive selection when it does occur. Barriers to dispersal, such as the Andes Mountains, can reduce the signature of hitchhiking by slowing the progression of sweeps (Barton 2000; Kim 2013). Indeed, at the two putative sweep loci investigated, similar but distinct alleles appear to have swept in the Eastern and Western populations. This is consistent with a scenario where a globally beneficial allele recombined onto a different genetic background as it spread, which would not only soften the sweep signal, but also enhance population differentiation (Slatkin and Wiehe 1998; Bierne 2010). The source of beneficial variation is another important factor, as adaptation from standing or introgressed variation, can result in soft sweeps (Pennings and Hermisson 2006). One likely example is the repeated evolution of certain wing pattern forms. Despite the strong selection known to act upon wing-patterning loci (Mallet and Barton 1989), signatures of selective sweeps at pattern loci have not been observed (Baxter *et al.* 2010; Nadeau *et al.* 2012). While the present study was not designed to address this question, all Western population samples shared a red forewing band, controlled by the *B* locus on chromosome 18 (Baxter *et al.* 2010); and yet no significant sweep signal was detected at this locus. It appears that wing patterning frequently evolves by sharing of preexisting alleles between populations, and even between species through rare hybridization (Pardo-Diaz *et al.* 2012; Heliconius Genome Consortium 2012; Wallbank *et al.* 2016). The presence of variation among these old alleles might eliminate any signature of genetic hitchhiking (Pennings and Hermisson 2006). It is yet to be established whether adaptation from standing and introgressed variation is generally common in *Heliconius*, but studies in other systems are increasingly suggesting an important role for preexisting adaptive variation in evolution (Jones *et al.* 2012; Gosset *et al.* 2014; Roesti *et al.* 2014).

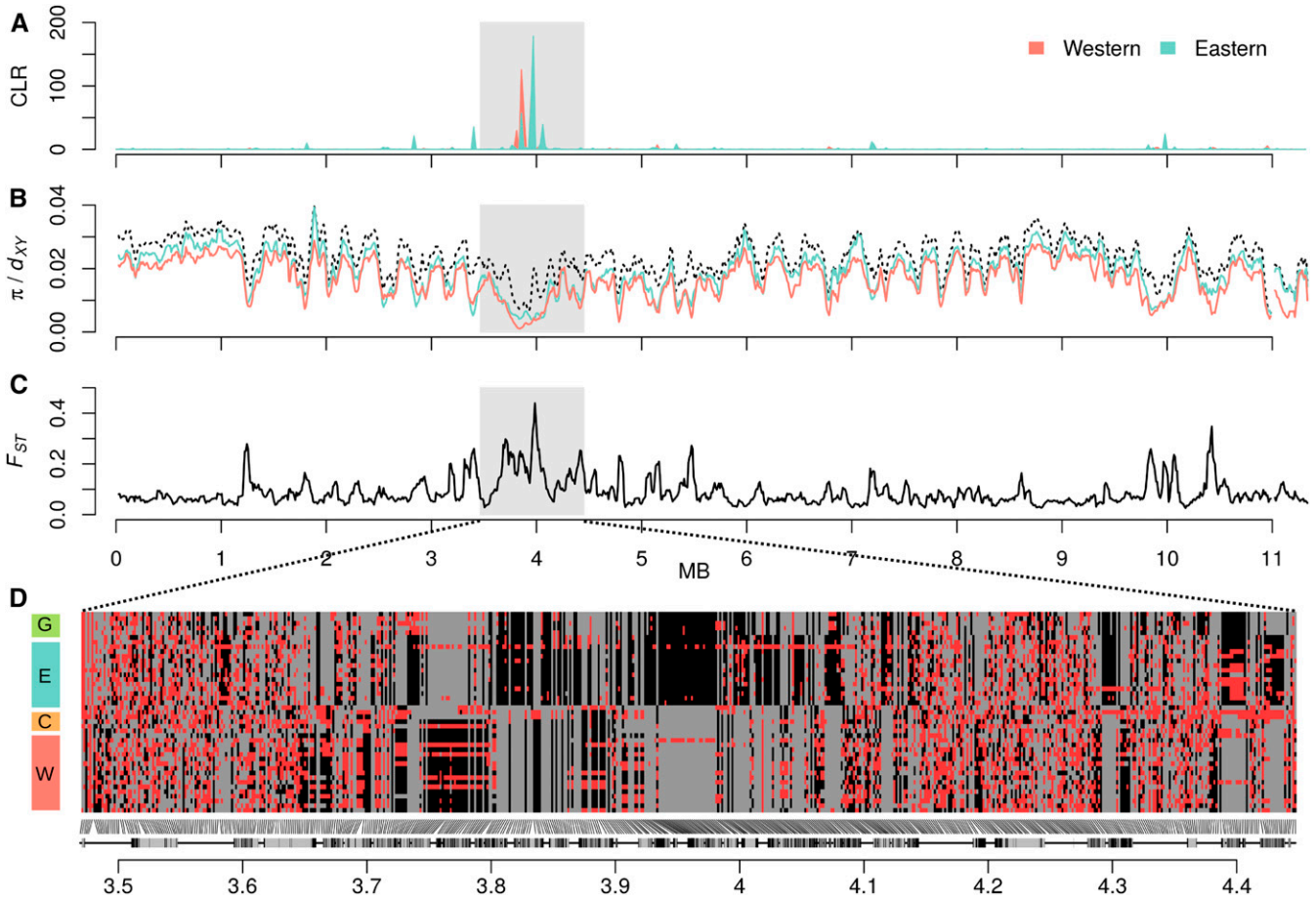

Figure 7 Putative selective sweep on chromosome 11. (A) Composite likelihood ratio (CLR) values calculated by SweeD (Pavlidis *et al.* 2013) for the Eastern and Western populations for 1000 windows across chromosome 11. Scaffolds are shaded light and dark. (B) Nucleotide diversity (π) for the Eastern and Western populations (in color) and divergence (d_{XY}) between these two populations (black dashed line), calculated for 50-kb sliding windows across chromosome 11, sliding in increments of 10 kb. (C) F_{ST} between the Eastern and Western populations was calculated in windows as in B. (D) Individual genotypes at 600 biallelic SNPs on scaffold HE672079, which harbors the putative selective sweep. Homozygous genotypes are colored gray (major allele) and black (minor allele), and heterozygotes are colored red. To optimize the detection of differences between populations, SNPs with a high degree of polymorphism (minor allele frequency ≥0.25) were considered. The 600 SNPs plotted were sampled semirandomly, ensuring that no two sampled SNPs were >1000 bp apart. Protein coding genes are indicated below the plot, with exons shown in black.

Despite the limited influence of hard selective sweeps, the genomic landscape of variation in *H. melpomene* has been shaped significantly by selection on linked sites. Diversity at neutral sites is positively correlated with recombination rate and negatively correlated with local gene density, which is a good proxy for the density of both coding and noncoding functional elements. These trends are consistent with a pervasive influence of selection on linked sites and are similar to patterns in several other animals (Cutter and Payseur 2013; Corbett-Detig *et al.* 2015). There is no evidence that recombination rate affects the mutation rate, in agreement with findings in *Drosophila* (McGaugh *et al.* 2012) and humans (McVicker *et al.* 2009). The effects of linked selection are also visible at the whole-chromosome scale. Longer chromosomes are less polymorphic, presumably because they have lower recombination rates per base pair, leading to stronger effects of linked selection on average. A negative relationship between chromosome length and recombination rate has been observed in a range of taxa (Kaback *et al.* 1992; Lander *et al.* 2001; Kawakami *et al.* 2014), and probably stems from a requirement for at least one obligate crossover event per chromosome during meiosis, even on the smallest chromosomes (Kawakami *et al.* 2014). Increased recombination rates on smaller chromosomes would also be expected to lead to more efficient purifying selection, due to reduced interference among loci (Hill and Robertson 1966; Felsenstein 1974). Indeed, smaller chromosomes display greater evidence of codon usage bias, a phenomenon probably driven by fairly weak selection. This is akin to the observation in *D. melanogaster* of reduced codon usage bias in regions of minimal recombination (Kliman and Hey 1993).

Further studies of other taxa are necessary to build a complete picture of how selection shapes genetic variation in natural populations. Nevertheless, even a simple comparison between butterflies and fruit flies can be enlightening. The different biology of these two insect groups results in

distinct patterns of adaptive evolution. However, the lower effective population sizes in *Heliconius* combined with less influence of selection on linked sites leads to levels of neutral diversity similar to those in *Drosophila* spp. Indeed, Corbett-Detig *et al.* (2015) estimate that the impact of selection on linked sites is around three times greater in fruit flies. Although it has long been recognized that levels of neutral variation are not solely determined by population size, whole genome studies such as this are beginning to reveal in detail how different processes combine to shape genetic diversity.

Acknowledgments

We thank the handling editor, David Begun, and two anonymous reviewers, whose comprehensive comments greatly improved this manuscript. Nick Barton and Aylwyn Scally provided helpful comments on an earlier draft. We thank John Davey for advice on the recombination rate analyses and for useful discussions of the results. James Walters provided advice for the analysis of the rate of adaptive substitution, and some of the groundwork for this analysis was performed by Gabriel Jamie and Joseph Harvey. We also thank James Mallet and Lawrence Gilbert for thought-provoking discussions of the life history of *Heliconius*. Finally, we thank Jenny Barna for computing support. This work was funded by European Research Council grants 281668 to F.M.J. and 339873 to C.D.J. Some sequencing was funded by a John Fell Fund of Oxford University grant to Judith Mank and James Walters. S.H.M. was supported by St. John's College, Cambridge. M.M. was supported by a Swiss National Science Foundation Early PostDoc Mobility fellowship (P2EZP3_148773) and an Austrian Science Fund Erwin Schrödinger fellowship (J3774-B25). C.S. was funded by Convocatoria para Proyectos de Investigación en Ciencias Básicas-2014 (Colciencias), contract no. FP44842-103-2015. W.J.P was supported by a Medical Research Council Centenary Award.

Literature Cited

Andolfatto, P., 2007 Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome. Genome Res. 17: 1755–1762.

Arias, C. F., C. Rosales, C. Salazar, J. Castaño, E. Bermingham *et al.*, 2012 Sharp genetic discontinuity across a unimodal *Heliconius* hybrid zone. Mol. Ecol. 21: 5778–5794.

Barton, N. H., 2000 Genetic hitchhiking. Philos. Trans. R. Soc. Lond. B Biol. Sci. 355: 1553–1562.

Barton, N., 2010 Understanding adaptation in large populations. PLoS Genet. 6: e1000987.

Baxter, S. W., N. J. Nadeau, L. S. Maroja, P. Wilkinson, B. A. Counterman *et al.*, 2010 Genomic hotspots for adaptation: the population genetics of Müllerian Mimicry in the *Heliconius* melpomene clade. PLoS Genet. **6:** e1000794.

Begun D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature **356:** 519–520.

Begun, D. J., A. K. Holloway, K. Stevens, L. W. Hillier, Y.-P. Poh *et al.*, 2007 Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biol. 5: e310.

Bierne, N., 2010 The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. Evolution 64: 3254–3272.

Campos, J. L., D. L. Halligan, P. R. Haddrill, and B. Charlesworth, 2014 The relation between recombination rate and patterns of molecular evolution and variation in drosophila melanogaster. Mol. Biol. Evol. 31: 1010–1028.

Charlesworth, B., 1998 Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. 15: 538–543.

Charlesworth, B., M. T. Morgan, and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289–1303.

Comeron, J. M., 2014 Background selection as baseline for nucleotide variation across the drosophila genome. PLoS Genet. 10: e1004434.

Corbett-Detig, R. B., D. L. Hartl, and T. B. Sackton, 2015 Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 13: e1002112.

Cutter, A. D., and A. M. Moses, 2011 Polymorphism, divergence, and the role of recombination in Saccharomyces cerevisiae genome evolution. Mol. Biol. Evol. 28: 1745–1754.

Cutter, A. D., B. A Payseur, 2013 Genomic signatures of selection at linked sites: unifying the disparity among species. Nat. Rev. Genet. 14: 262–274.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–498.

Drummond, D. A., A. Raval, and C. O. Wilke, 2006 A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. 23: 327–337.

Durbin, J., and G. S. Watson, 1950 Testing for serial correlation in least squares regression. I. Biometrika 37: 409–428.

Durbin, J., and G. S. Watson, 1951 Testing for serial correlation in least squares regression. II. Biometrika 38: 159–178.

Ehrlich, P. R., and L. E. Gilbert, 1973 Population structure and dynamics of the tropical butterfly *Heliconius* ethilla. Biotropica 5: 69–82.

Enard, D., P. W. Messer, and D. Petrov a, 2014 Genome-wide signals of positive selection in human evolution. Genome Res. 24: 885–895.

Eyre-Walker, A., 2006 The genomic rate of adaptive evolution. Trends Ecol. Evol. 21: 569–575.

Eyre-Walker, A., and P. D. Keightley, 2009 Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol. Biol. Evol. 26: 2097–2108.

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.

Felsenstein, J., 1974 The evolution advantage of recombination. Genetics 78: 737–756.

Field, A., 2009 Discovering Statistics Using SPSS, SAGE Publications, London.

Fu, Y. X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. 48: 172–197.

Gillespie, J. H., 2001 Is the population size of a species relevant to its evolution? Evolution 55: 2161–2169.

Gosset, C. C., J. Do Nascimento, M.-T. Augé, and N. Bierne, 2014 Evidence for adaptation from standing genetic variation on an antimicrobial peptide gene in the mussel Mytilus edulis. Mol. Ecol. 23: 3000–3012.

Heliconius Genome Consortium, 2012 Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487: 94–8.

Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton *et al.*, 2011 Classic selective sweeps were rare in recent human evolution. Science 331: 920–924.
Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. 8: 269–294.
Huang, W., L. Li, J. R. Myers, and G. T. Marth, 2012 ART: a next-generation sequencing read simulator. Bioinformatics 28: 593–594.
Hudson, R. R., D. Boos, and N. Kaplan, 1992 A statistical test for detecting geographic subdivision. Mol. Biol. Evol. 9: 138–151.
Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli *et al.*, 2012 The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484: 55–61.
Kaback D., V. Guacci, D. Barber, J. Mahon, 1992 Chromosome size-dependent control of meiotic recombination. Science 256: 228–232.
Kawakami, T., L. Smeds, N. Backström, A. Husby, A. Qvarnström *et al.*, 2014 A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. Mol. Ecol. 23: 4035–4058.
Keightley, P. D., A. Pinharanda, R. W. Ness, F. Simpson, K. K. Dasmahapatra *et al.*, 2014 Estimation of the spontaneous mutation rate in *Heliconius* melpomene. Mol. Biol. Evol. 32: 239–243.
Kim, Y., 2013 Stochastic patterns of polymorphism after a selective sweep over a subdivided population. Genet. Res. 95: 57–67.
Kliman, R., and J. Hey, 1993 Reduced natural selection associated with low recombination in Drosophila melanogaster. Mol. Biol. Evol. 10: 1239–1258.
Kozak, K. M., N. Wahlberg, A. Neild, K. K. Dasmahapatra, J. Mallet *et al.*, 2015 Multilocus species trees show the recent adaptive radiation of the mimetic. Syst. Biol. 64: 505–524.
Kronforst, M. R. R., M. E. B. Hansen, N. G. G. Crawford, J. R. R. Gallant, W. Zhang *et al.*, 2013 Hybridization reveals the evolving genomic architecture of speciation. Cell Reports 5: 666–677.
Lander, E. S., L. M Linton, B Birren, C Nusbaum, M. C Zody, 2001 Initial sequencing and analysis of the human genome. Nature 409: 860–921.
Langley, C. H., K Stevens, C Cardeno, Y. C Lee, D. R Schrider, 2012 Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192: 533–598.
Lee, Y. C. G., C. H. Langley, and D. J. Begun, 2014 Differential strengths of positive selection revealed by hitchhiking effects at small physical scales in drosophila melanogaster. Mol. Biol. Evol. 31: 804–816.
Leffler, E. M., K. Bullaughey, D. R. Matute, W. K. Meyer, L. Ségurel *et al.*, 2012 Revisiting an old riddle: What determines genetic diversity levels within species? PLoS Biol. 10: e1001388.
Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change*, Columbia University Press, New York.
Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.
Li, J., H. Li, M. Jakobsson, S. Li, P. Sjödin *et al.*, 2012 Joint analysis of demography and selection in population genetics: Where do we stand and where could we go? Mol. Ecol. 21: 28–44.
Lohmueller, K. E., A. Albrechtsen, Y. Li, S. Y. Kim, T. Korneliussen *et al.*, 2011 Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet. 7: e1002326.
Lunter, G., and M. Goodson, 2011 Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21: 936–939.
Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles *et al.*, 2012 The Drosophila melanogaster Genetic Reference Panel. Nature 482: 173–178.
Mallet, J., and N. Barton, 1989 Strong natural selection in a warning-color hybrid zone. Evolution (N. Y.) 43: 421–431.
Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters *et al.*, 2013 Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. Genome Res. 23: 1817–1828.
Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23.
McGaugh, S. E., C. S Heil, B Manzano-Winkler, L Loewe, S Goldstein, 2012 Recombination modulates how selection affects linked sites in Drosophila. PLoS Biol. 10: e1001422.
McVicker, G., D. Gordon, C. Davis, and P. Green, 2009 Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 5: e1000471.
Merrill, R. M., K. K Dasmahapatra, J. W Davey, D. D Dell'Aglio, J. J Hanly, 2015 The diversification of *Heliconius* butterflies: What have we learned in 150 years? J. Evol. Biol. 28: 1417–1438.
Messer, P. W., and D. a. Petrov, 2013 Frequent adaptation and the McDonald-Kreitman test. Proc. Natl. Acad. Sci. USA 110: 8615–8620.
Mevik, B. H., R. Wehrens, 2007 The pls Package: Principle Component and Partial Least Squares Regression in R. J. Stat. Softw. 18: 1–24.
De Mita, S., and M. Siol, 2012 EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genet. 13: 27.
Mugal, C. F., B. Nabholz, and H. Ellegren, 2013 Genome-wide analysis in chicken reveals that local levels of genetic diversity are mainly governed by the rate of recombination. BMC Genomics 14: 86.
Nachman, M. W., and B. A Payseur, 2012 Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Philos. Trans. R. Soc. Lond. B Biol. Sci. 367: 409–421.
Nachman, M. W., V. L. Bauer, S. L. Crowell, and C. F. Aquadro, 1998 DNA variability and recombination rates at X-linked loci in humans. Genetics 150: 1133–1141.
Nadeau, N. J., A. Whibley, R. T. Jones, J. W. Davey, K. K. Dasmahapatra *et al.*, 2012 Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. Philos. Trans. R. Soc. B Biol. Sci. 367: 343–353.
Nadeau, N. J., S. H. Martin, K. M. Kozak, C. Salazar, K. K. Dasmahapatra *et al.*, 2013 Genome-wide patterns of divergence and gene flow across a butterfly radiation. Mol. Ecol. 22: 814–826.
Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res. 15: 1566–1575.
Ohta, T., and J. Gillespie, 1996 Development of neutral and nearly neutral theories. Theor. Popul. Biol. 49: 128–142.
Ott, M., J. Zola, A. Stamatakis, and S. Aluru, 2007 Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. Proceedings of the 2007 ACM/IEEE Conference on Supercomputing, Reno, NV.
Pardo-Diaz, C., C. Salazar, S. W. Baxter, C. Merot, W. Figueiredo-Ready *et al.*, 2012 Adaptive introgression across species boundaries in *Heliconius* butterflies. PLoS Genet. 8: e1002752.
Pavlidis, P., D. Živkovic, A. Stamatakis, and N. Alachiotis, 2013 SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol. Biol. Evol. 30: 2224–2234.
Payseur, B. A., and M. W. Nachman, 2002 Gene density and human nucleotide polymorphism. Mol. Biol. Evol. 19: 336–340.
Pennings, P. S., and J. Hermisson, 2006 Soft sweeps II: molecular population genetics of adaptation from recurrent mutation or migration. Mol. Biol. Evol. 23: 1076–1084.
Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38: 904–909.
Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Rambaut, A., and N. C. Grass, 1997 Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics 13: 235–238.
Roesti, M., S. Gavrilets, A. P. Hendry, W. Salzburger, and D. Berner, 2014 The genomic signature of parallel adaptation from shared genetic variation. Mol. Ecol. 23: 3944–3956.
Sattath, S., E. Elyashiv, O. Kolodny, Y. Rinott, and G. Sella, 2011 Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in Drosophila simulans. PLoS Genet. 7: e1001302.
Sella, G., D. A Petrov, M. Przeworski, and P. Andolfatto, 2009 Pervasive natural selection in the Drosophila genome? PLoS Genet. 5: e1000495.
Slatkin, M., and T. Wiehe, 1998 Genetic hitch-hiking in a subdivided population. Genet. Res. 71: 155–160.
Stamatakis, A., 2006 RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.
Stamatakis, A., P. Hoover, and J. Rougemont, 2008 A rapid bootstrap algorithm for the RAxML Web servers. Syst. Biol. 57: 758–771.
Supple, M. A., H. M. Hines, K. K. Dasmahapatra, J. J. Lewis, D. M. Nielsen *et al*., 2013 Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies. Genome Res. 23: 1248–1257.
Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.
Wallbank, R. W. R., S. W. Baxter, C. Pardo-Díaz, J. J. Hanly, S. H. Martin *et al.*, 2016 Evolutionary novelty in a butterfly wing pattern through enhancer shuffling. PLoS Biol. 14: e1002353.
Welch, J. J., 2006 Estimating the genomewide rate of adaptive protein evolution in Drosophila. Genetics 173: 821–837.
Wright, F., 1990 The "effective number of codons" used in a gene. Gene 87: 23–29.

Communicating editor: D. J. Begun

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.183285/-/DC1

Natural Selection and Genetic Diversity in the Butterfly *Heliconius melpomene*

Simon H. Martin, Markus Möst, William J. Palmer, Camilo Salazar, W. Owen McMillan, Francis M. Jiggins, and Chris D. Jiggins

DOI: 10.1534/genetics.115.183285

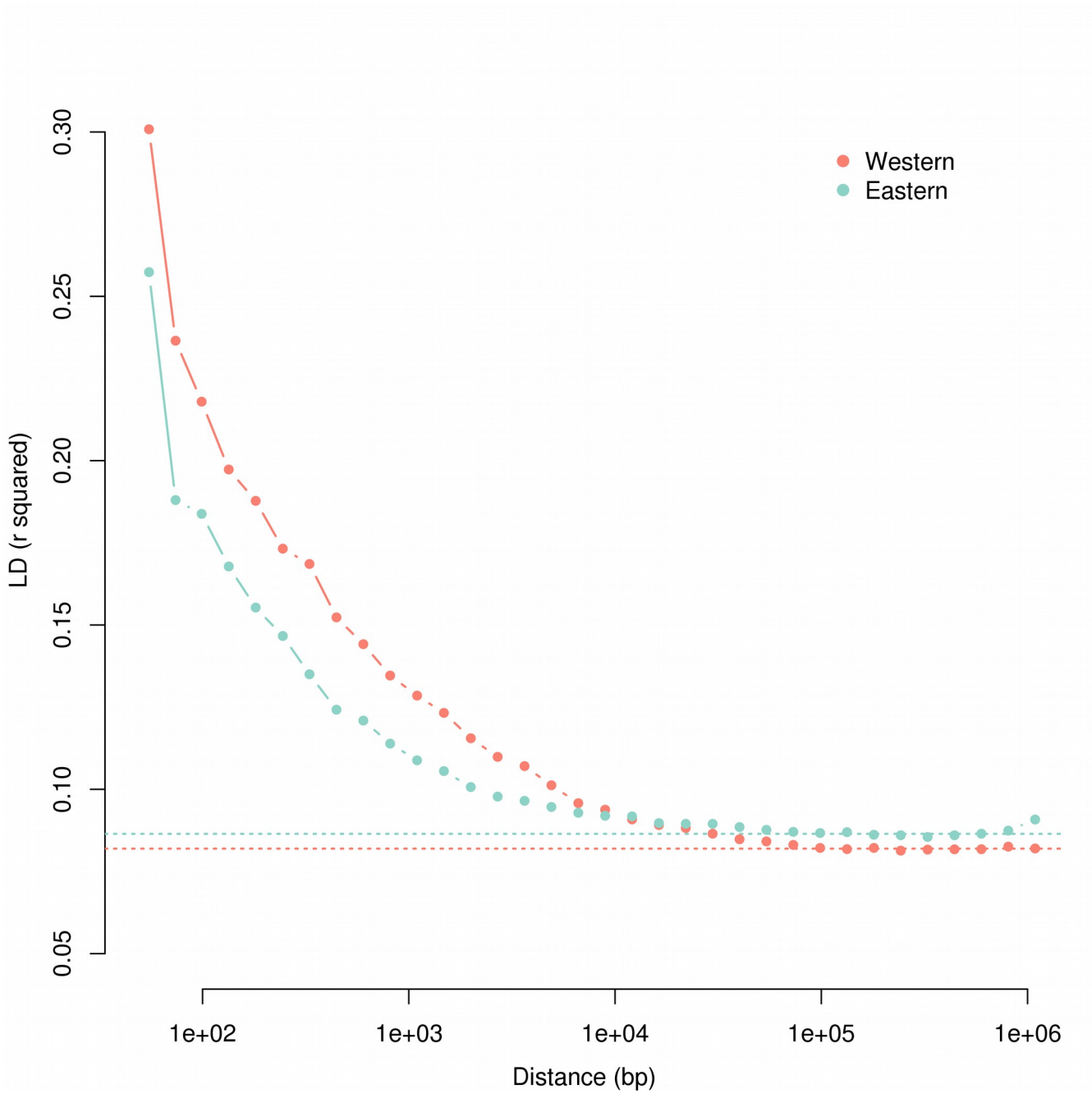

Fig. S1 Linkage disequilibrium (r^2) plotted by distance between SNP pairs

LD was calculated for all SNP pairs with minor allele counts of at least 5 on the top 100 largest scaffolds. SNP pairs were binned by distance in bins of logarithmically increasing size. Dashed lines indicate background r^2, calculated between unlinked SNPs on different chromosomes.

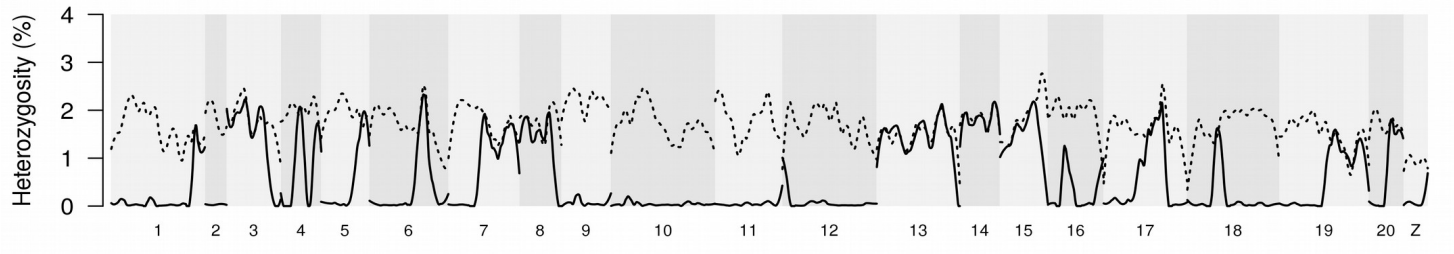

Fig. S2 Heterozygosity in inbred and outbred individuals

Proportion of heterozygous genotypes in the inbred reference genome individual, melP.HGC1 (solid line), and from a wild-caught sample from the same source population in Panama, melP.CJ18097 (dashed). Lines are smoothed using loess (local regression), for each chromosome with a span equivalent to 3 Mb. The 21 chromosomes are shaded. Each chromosome was constructed by concatenating scaffolds according to their placement in the *H. melpomene* genome v1.1 linkage map.

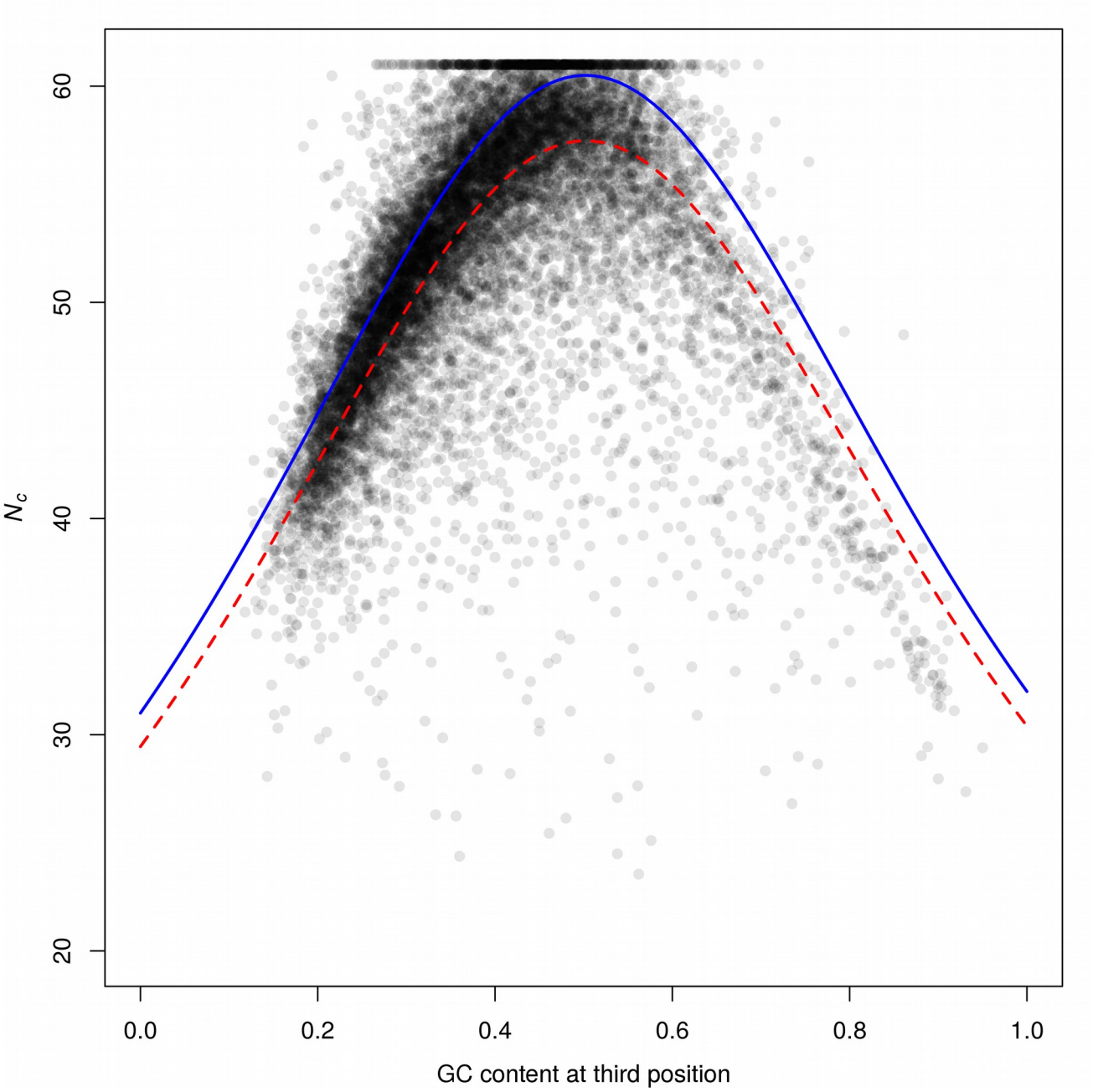

Fig. S3 Codon Usage

Each point represents a gene. The effective number of codons (Nc) is plotted against GC content at the third codon position. The expected value for Nc in the absence of codon usage bias is given by the blue line (Wright 1990). Points below the line indicate genes with lower codon variability than expected without codon usage bias. The red line indicates 95% of the null expectation, which formed our cut-off for classifying a gense as having low codon usage bias.

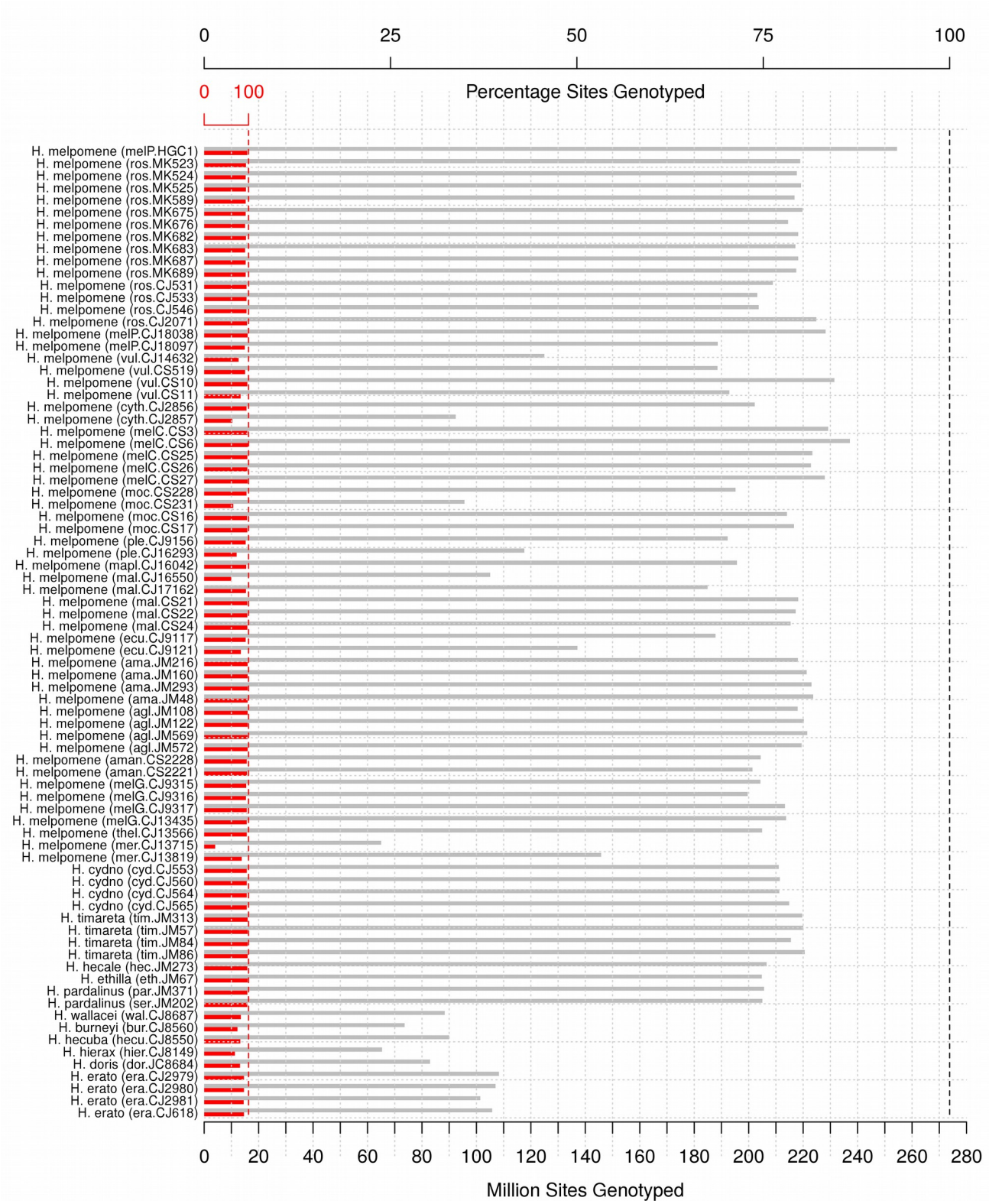

Fig. S4 Number of genotyped sites per individual

Horizontal bars show the number of sites (bottom axis) with high-quality genotypes per individual, considering either all sites (grey) or only coding sites (red). The top axis shows the percentage of sites genotyped, out of 273 Mb for the whole genome (black), or the 16.3 Mb of coding sites (red). Although the more distantly related species have far fewer sites genotyped overall, they have a good proportion of coding sites genotyped. Some individuals have low numbers of genotypes called overall and among coding sites, suggesting that these are low-coverage or low-quality samples.

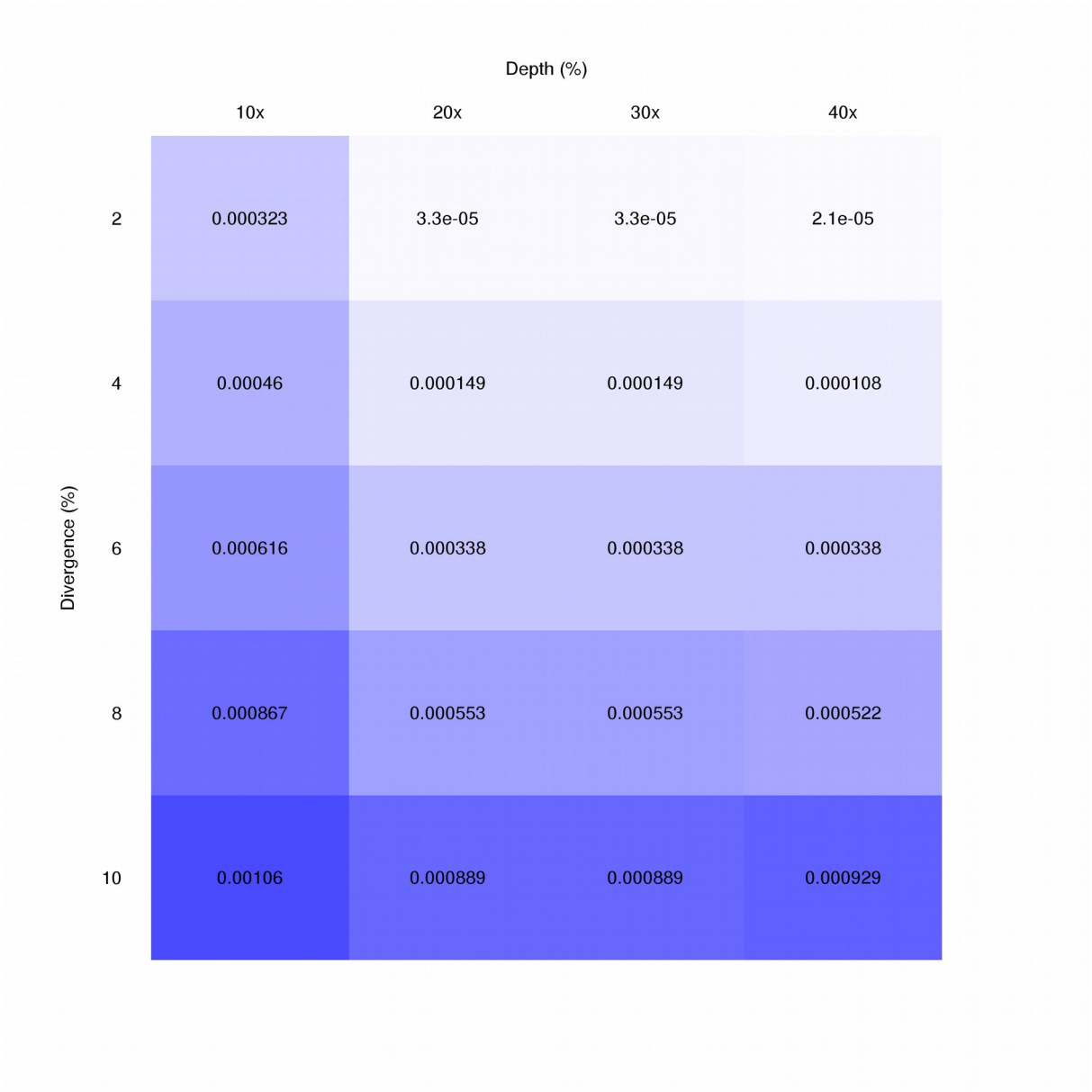

Fig. S5. Estimated error rates based on simulated sequence data

The number of incorrect genotype calls per site over 1 million sites.

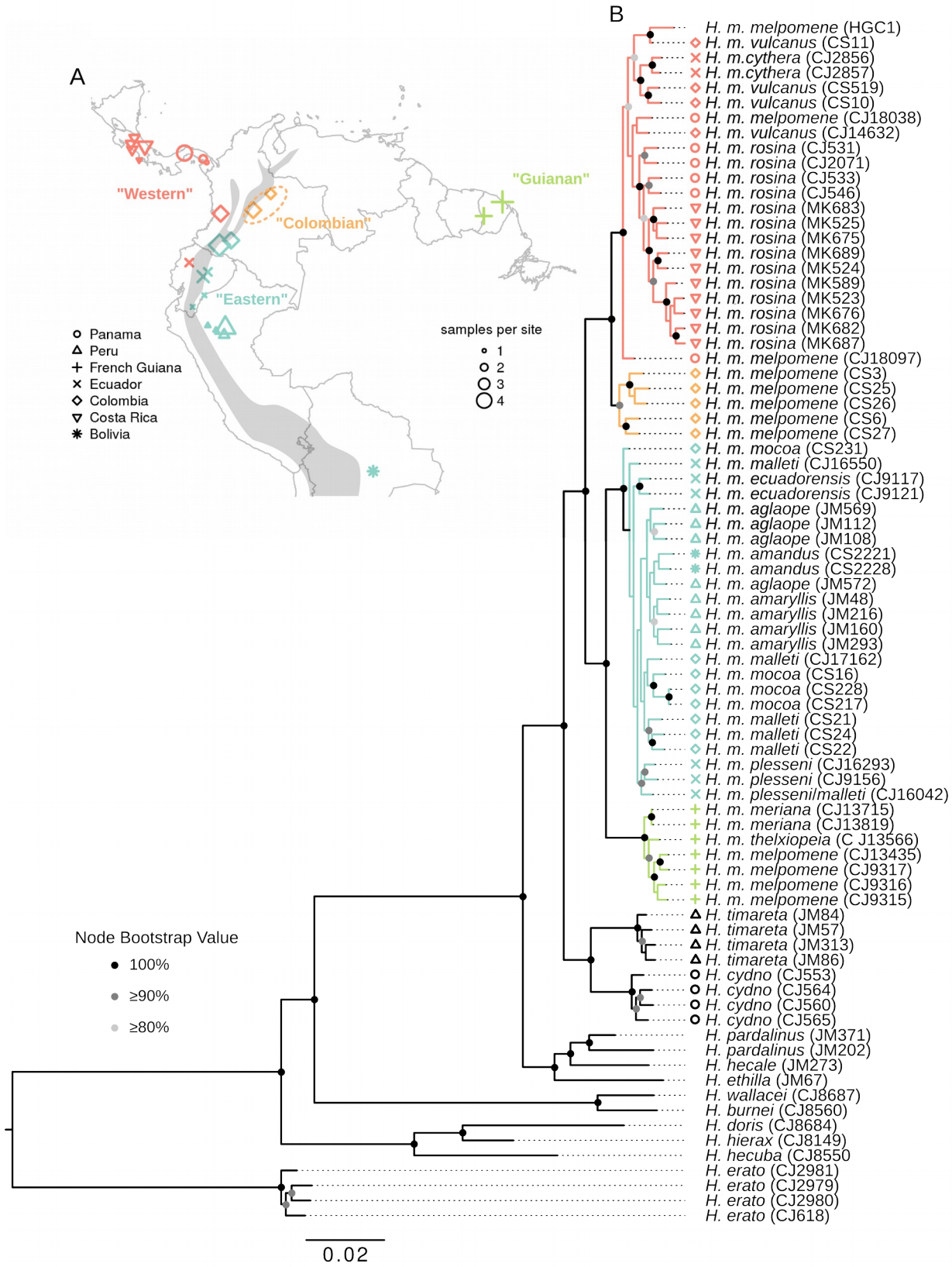

Fig. S6 Whole genome ML tree

A. Sampling locations of the 58 wild *H. melpomene* samples (see Table S1 for coordinates). Symbols indicate country of sampling, sizes indicate the number of samples from a location. Colours correspond to the four major populations as described in the main text. Grey shading indicates the approximate location of the Andes mountains. B. RaxML phylogeny based on four-fold degenerate

(4D) sites. Nodes with bootstrap support of at least 80%, 90% or 100% are indicated (see legend). Colours and symbols are as in A.

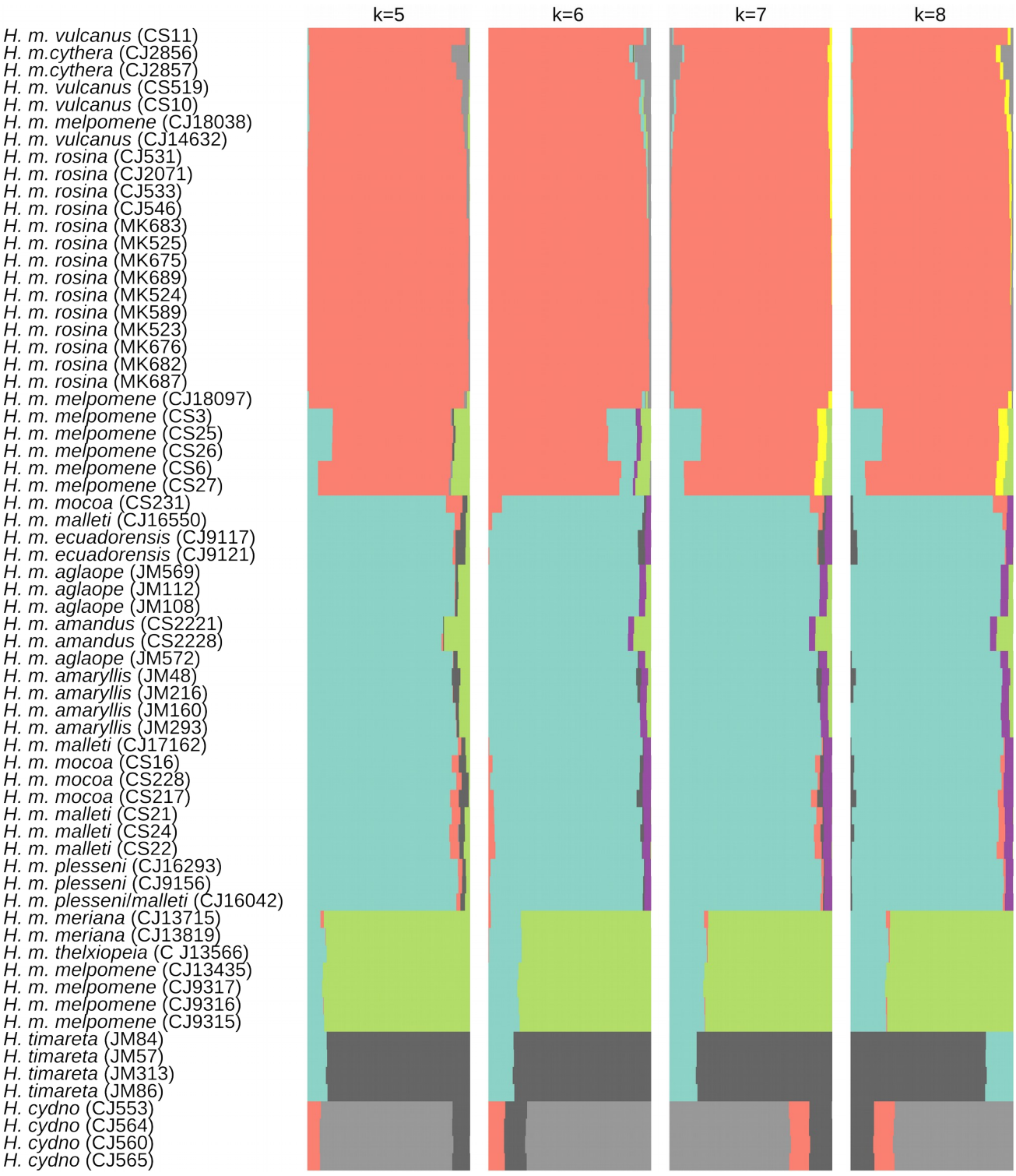

Fig. S7 Cluster assignments inferred using STRUCTURE

The runs with the highest probability out of five replicates for each k value are shown (k=3 and k=4 are not shown as they gave low probabilities). Each individual is assigned a probability of falling into each of the k clusters, as indicated by colours.

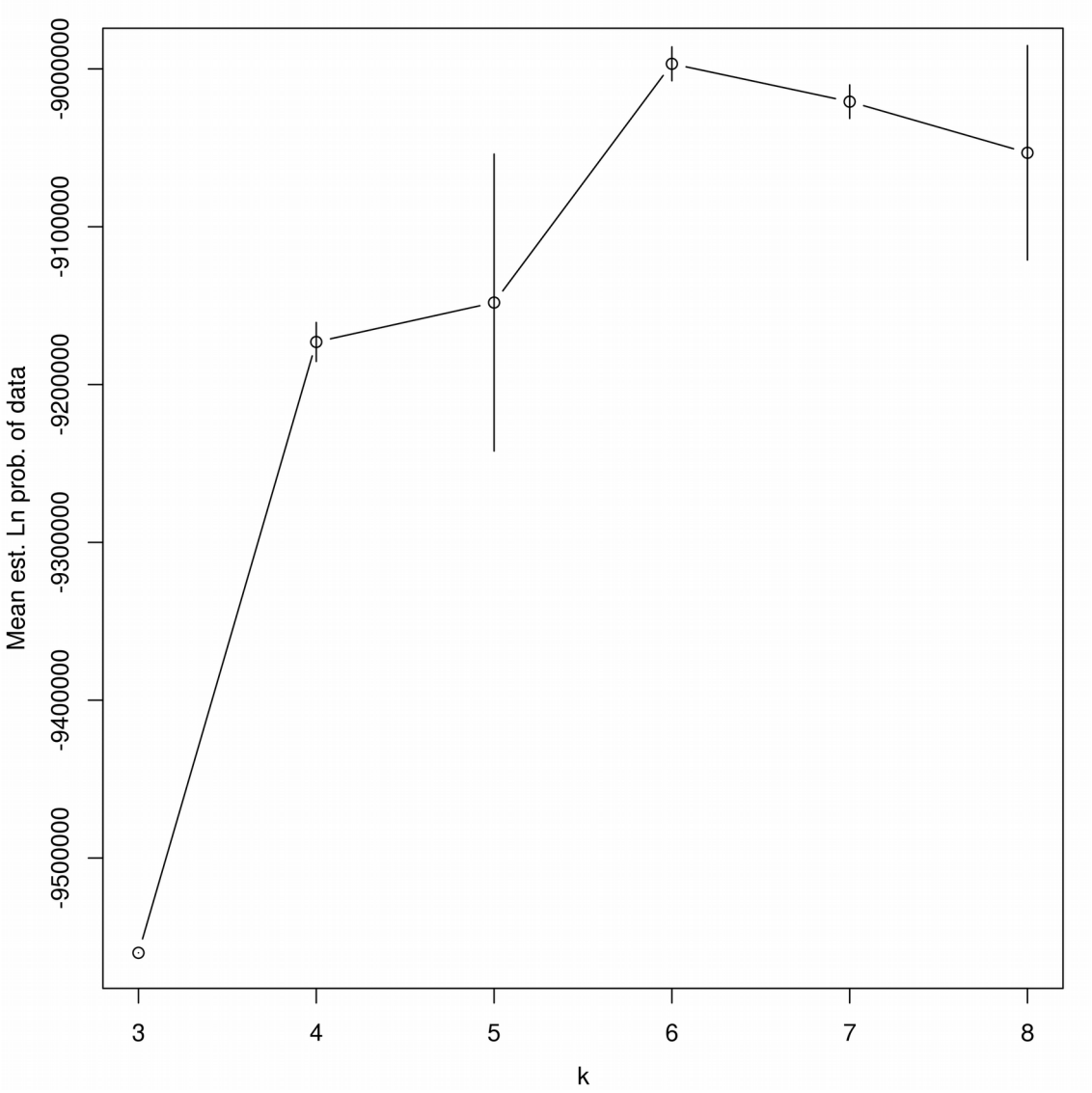

Fig. S8 Mean estimated Ln Probabilities of the data according to STRUCTURE

Five runs for each value of k from 3 to 8 were executed in STRUCTURE. The estimated mean and standard deviation of the Ln Likelihoods for each value of k, as estimated using STRUCTURE HARVESTER (Earl and vonHoldt 2012), are shown.

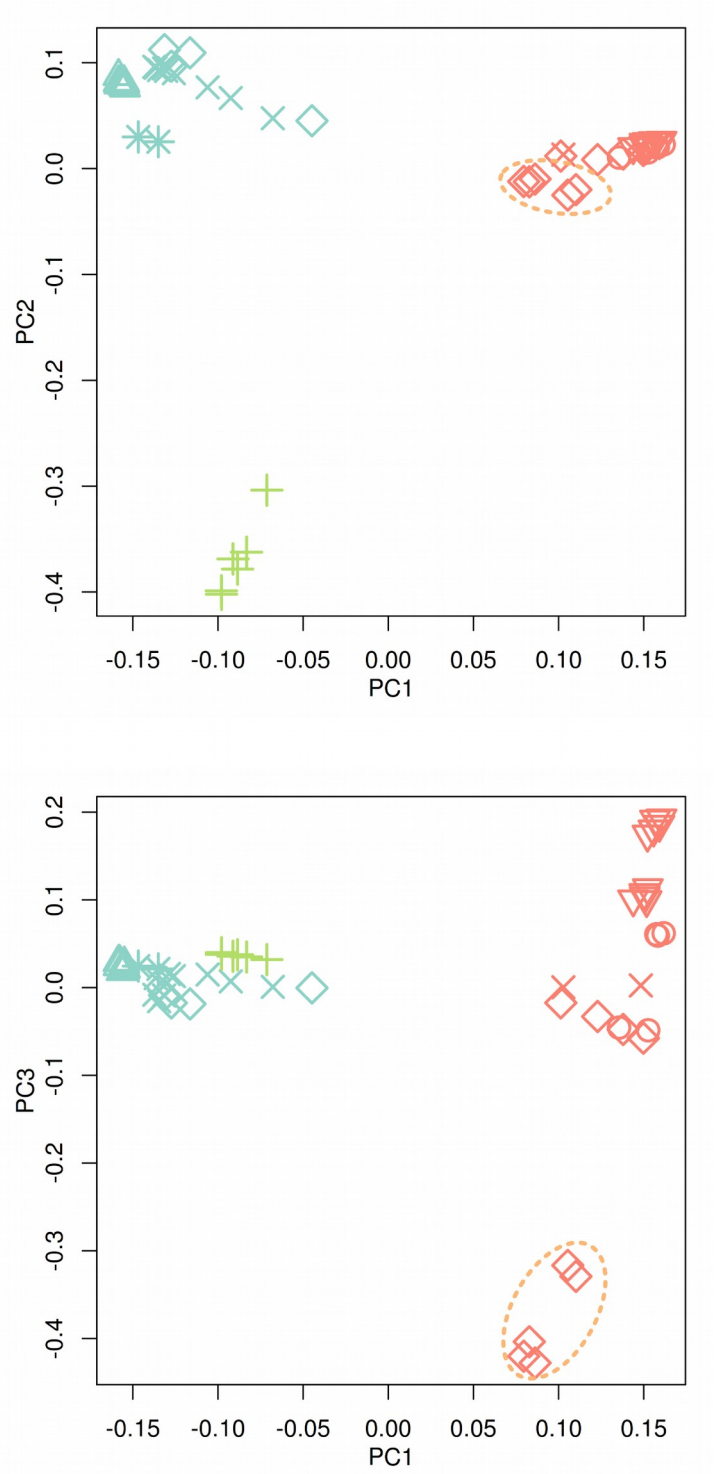

Fig. S9 Principal Components 1 & 2 and 1 & 3

Principal Component 1 (explained 17% of the variance) plotted against Principal Components 2 (7.7%) and 3 (3.4%). Colours as in Fig. 1C. The Colombian samples discussed in the main text are circled in orange.

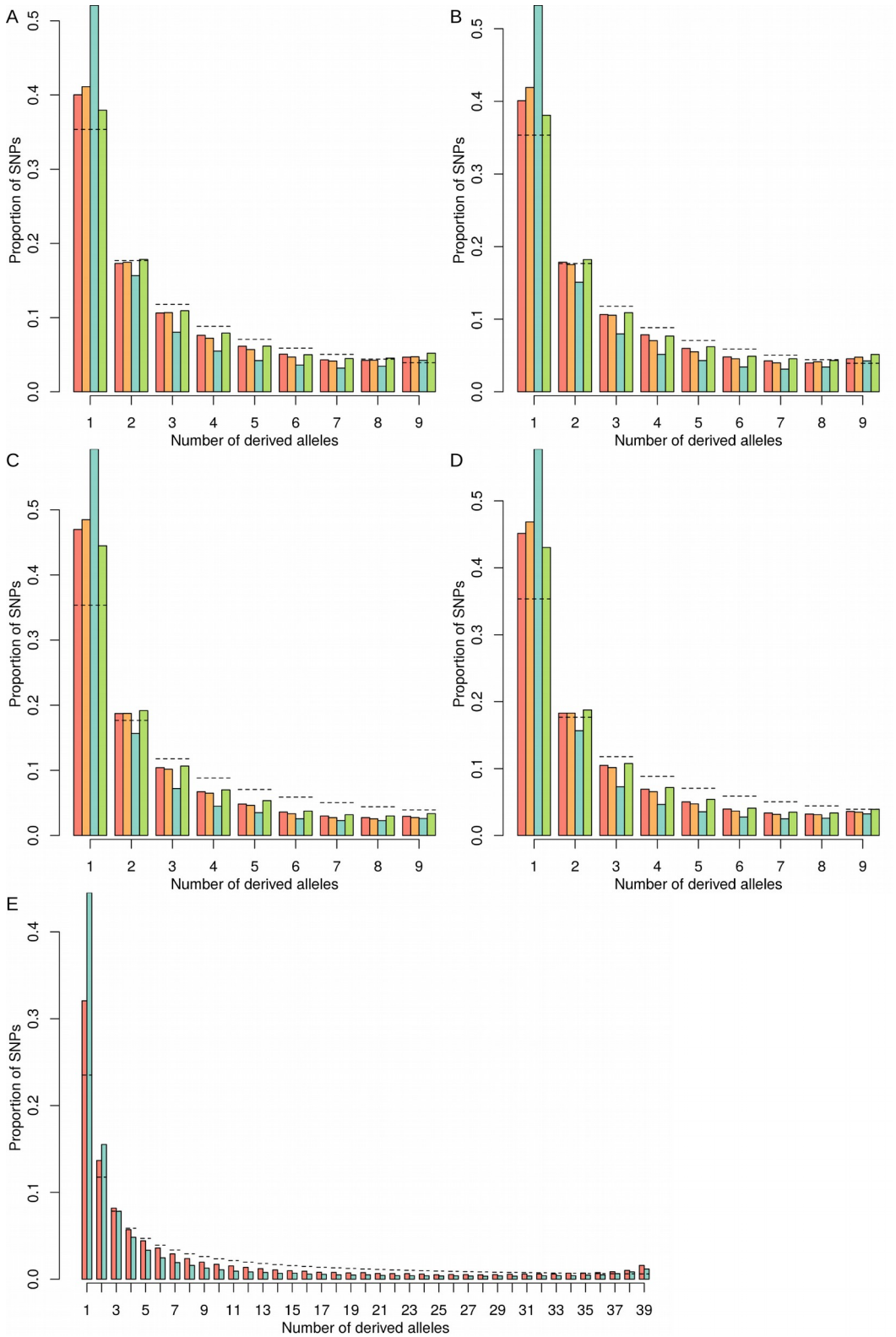

Fig. S10 Unfolded site frequency spectra for different sample sets and site classes

A. SFS for 4D sites, down-sampling to five samples per site. B. SFS for 4D sites, using only five high-depth samples from the same approximate sampling location for each population. C. SFS for intergenic sites, down-sampling to five samples per site. D. SFS for intronic sites, down-sampling to five samples per site. E. SFS for 4D sites, down-sampling to twenty samples per site. Populations are coloured are as follows, Western: red, Colombian: orange, Eastern: blue, Guianan: green. Dashed lines indicate the expected frequencies under the standard coalescent model with constant population size (Fu 1995).

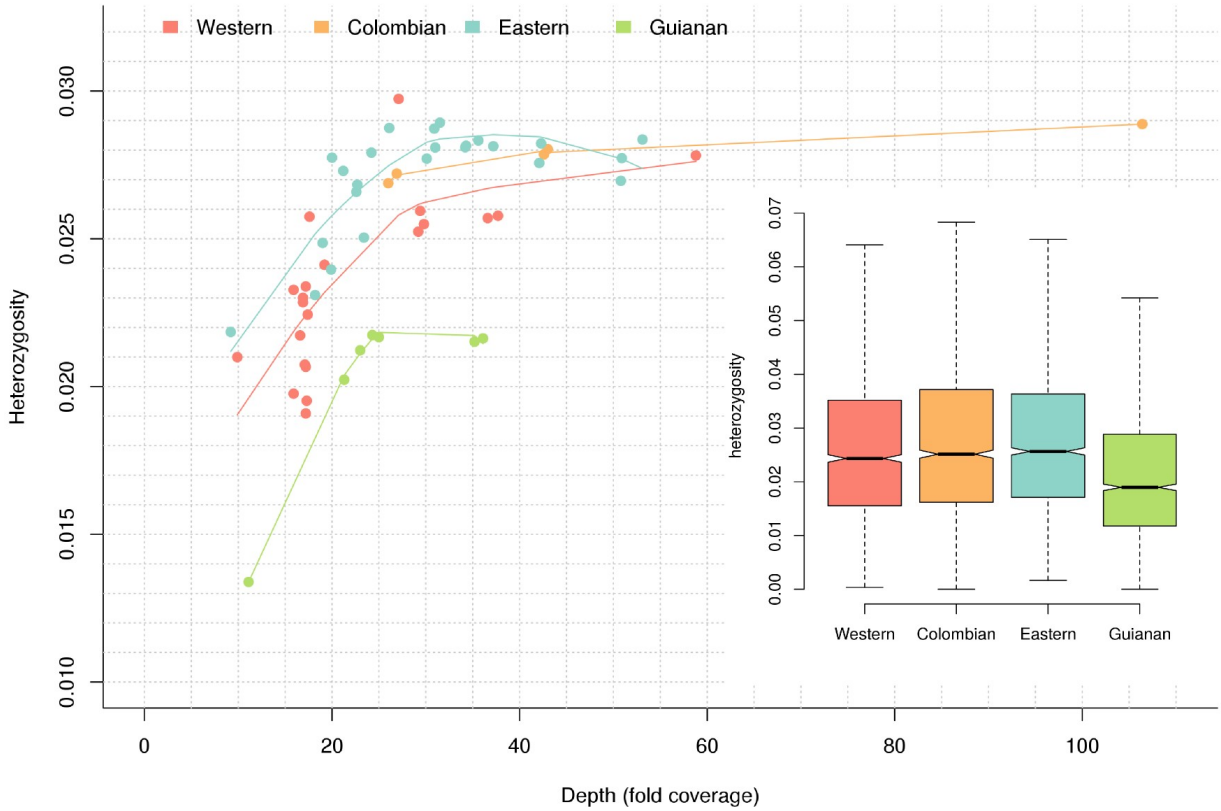

Fig. S11 Effect of sequencing depth on heterozygosity

Mean individual heterozygosity plotted against sequencing depth (fold coverage). Dashed lines indicate local regression smother, with span=1. Populations are plotted separately. **Insert:** boxplots of heterozygosity for all 100kb windows for each population, averaged over all samples with sequencing depth of at least 25x.

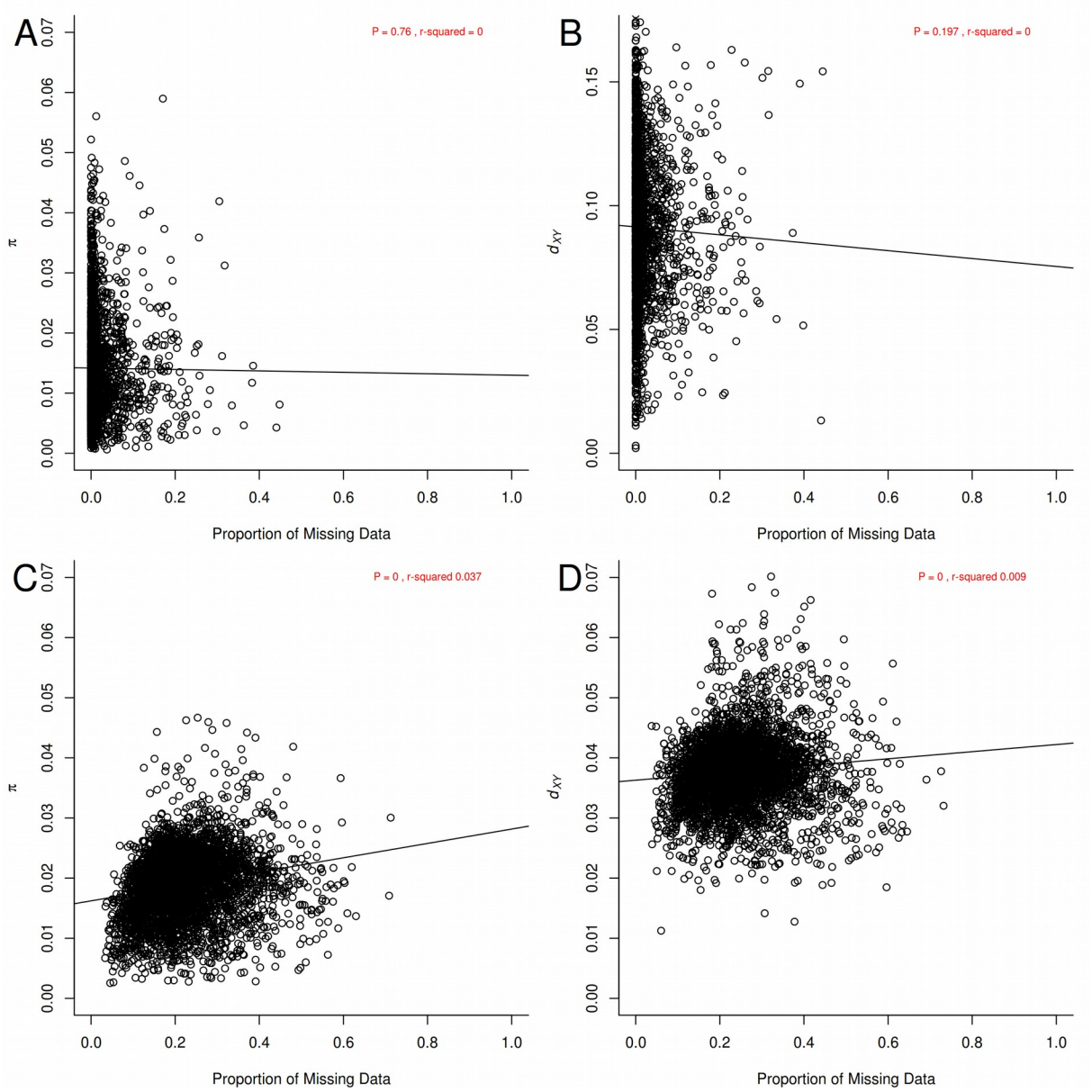

Fig. S12 Effects of missing data on measures of diversity and divergence

A. Diversity (π) in *H. melpomene* at third codon positions in 100 kb windows, plotted against the proportion of missing data per window (number of third codon positions that were genotyped in fewer than 50% of individuals). **B.** As in A, except divergence (d_{XY}) between *H. melpomene* and *H. erato* is plotted against the proportion of missing data. **C.** As in A, except using intergenic sites. **D.** as in B, except using intergenic sites, and d_{XY} is measured between *H. melpomene* and the silvaniform species, as *H. erato* was too divergent at intergenic sites. In all plots, the slope of a linear regression is shown, along with the P-value and r^2 for the linear model.

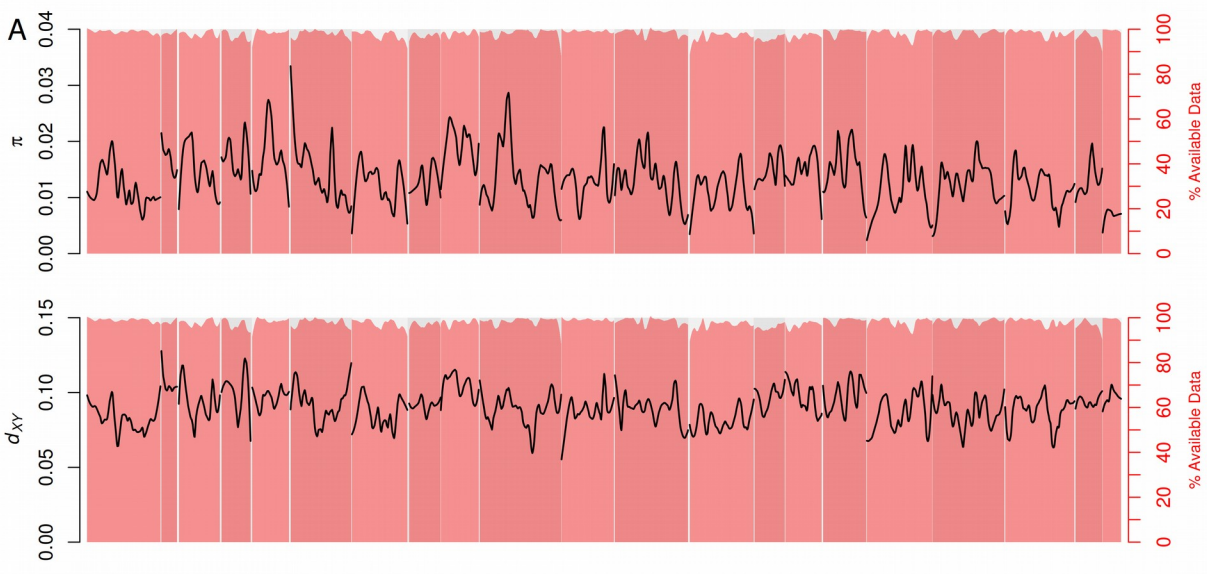

Fig. S13. Relationship between missing data and π and d_{XY}

Diversity (π) at third codon positions in *H. melpomene* and divergence (d_{XY}) at third codon positions between *H. melpomene* and *H. erato*, for 100 kb windows, smoothed using loess with a span equivalent to 3 Mb. Coloured fill indicates the percentage of available data per window (third codon positions gentotyped in at least half of the samples in the complete dataset).

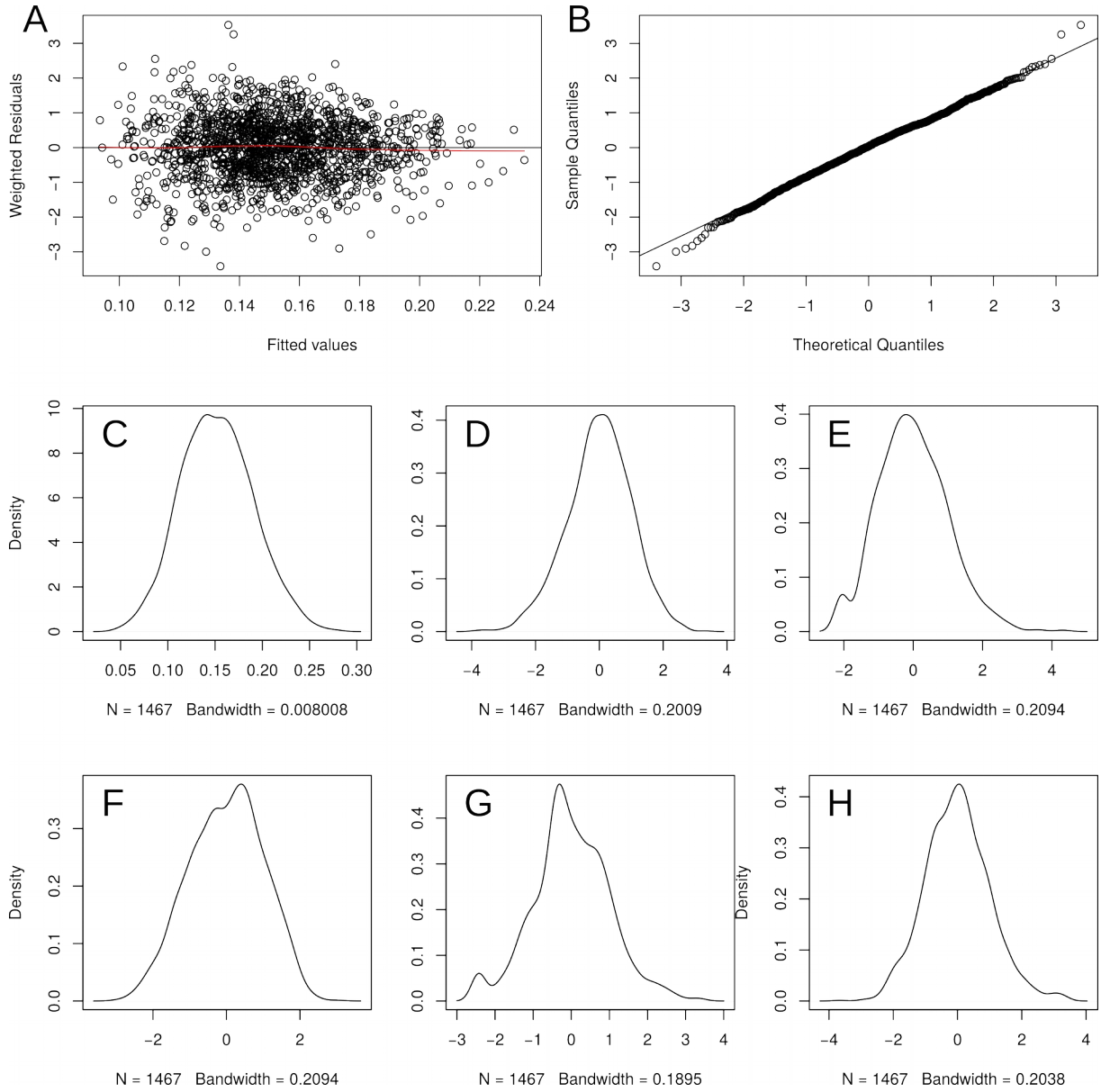

Fig. S14 Multiple regression residual plots and variable distributions

A. Weighted residuals of the main model plotted against fitted values. **B.** Normal QQ plot. **C-H.** Density plots for Z-transformed variables, C: π_{4D}, D: d_S, E: D_n, F: gene density, G: $\hat{r}$, H: GC-content.

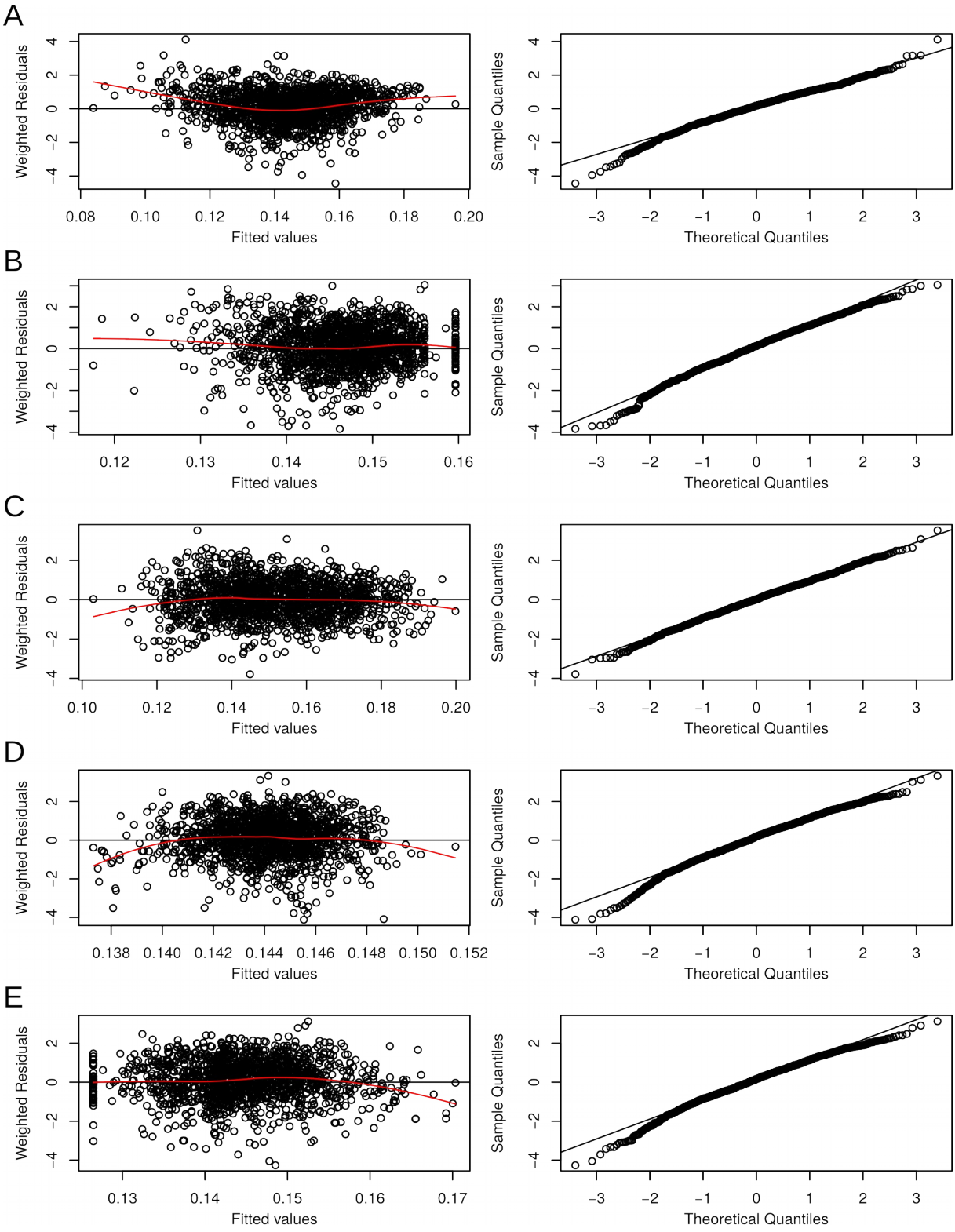

Fig. S15. Residual plots and QQ plots for explanatory variables of the main model

A-D. Weighted residual plots for all predictor variables of the main model (left) and normal QQ plots (right). A: d_S, B: D_n, C: gene density, D: GC-content, E: $\hat{r}$.

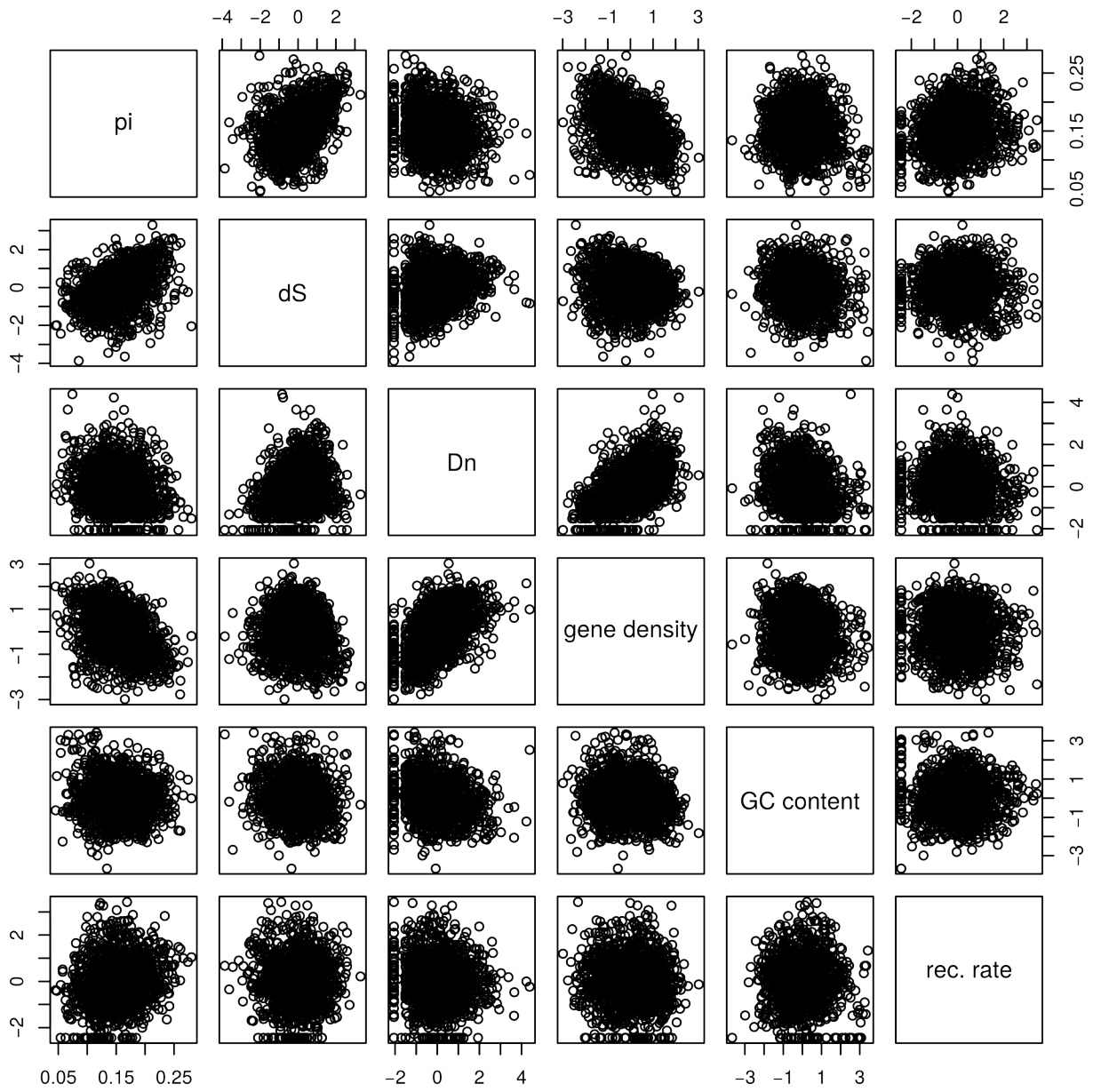

Fig. S16. Pairwise scatter plots of Z-transformed variables from the multiple regression main model.

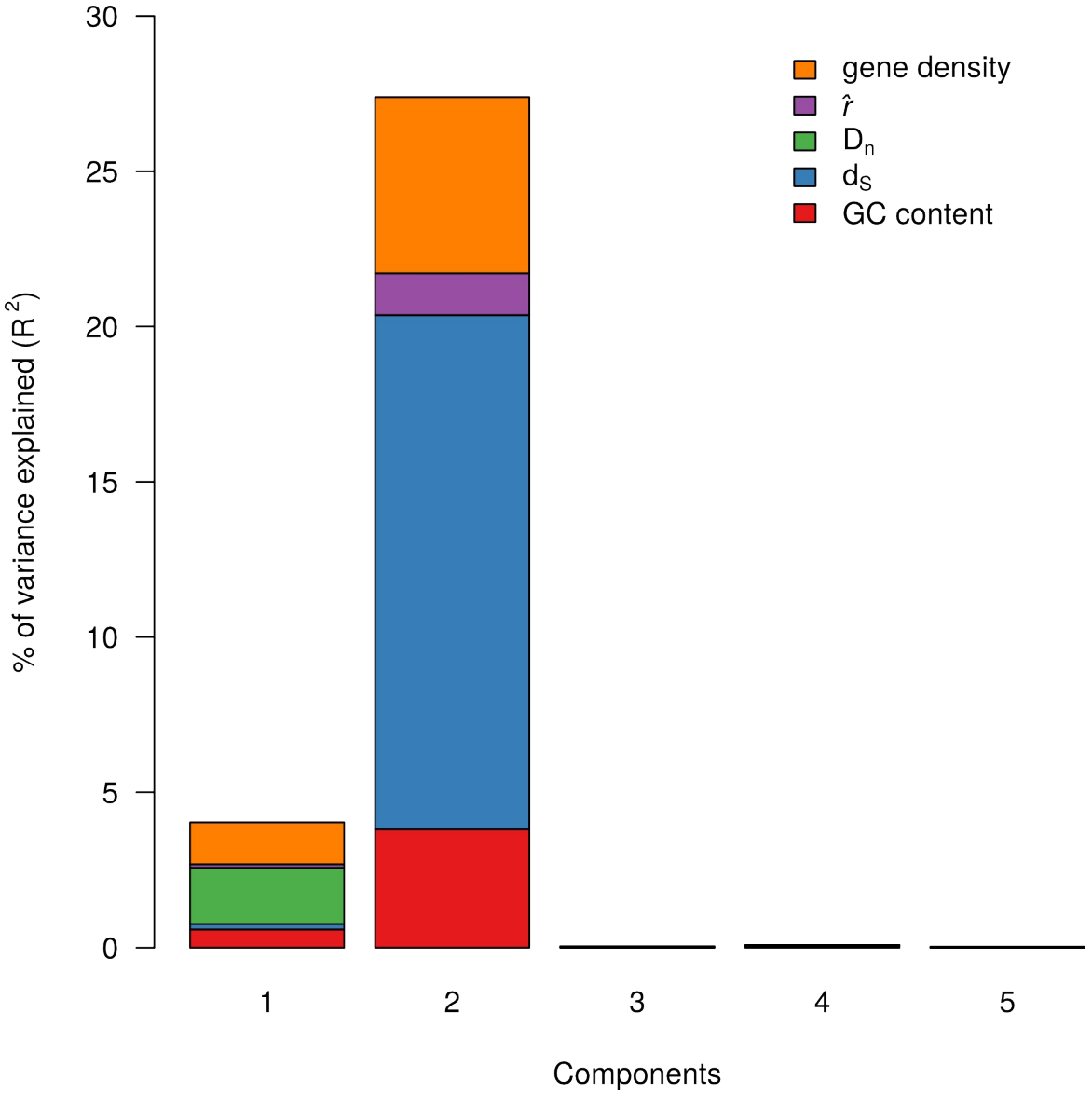

Fig. S17. Principle Component Regression (PCR)

Decomposition of effects of explanatory variables using PCR. Bars indicate the proportion of variance explained by each component, with colours indicating the five explanatory variables.

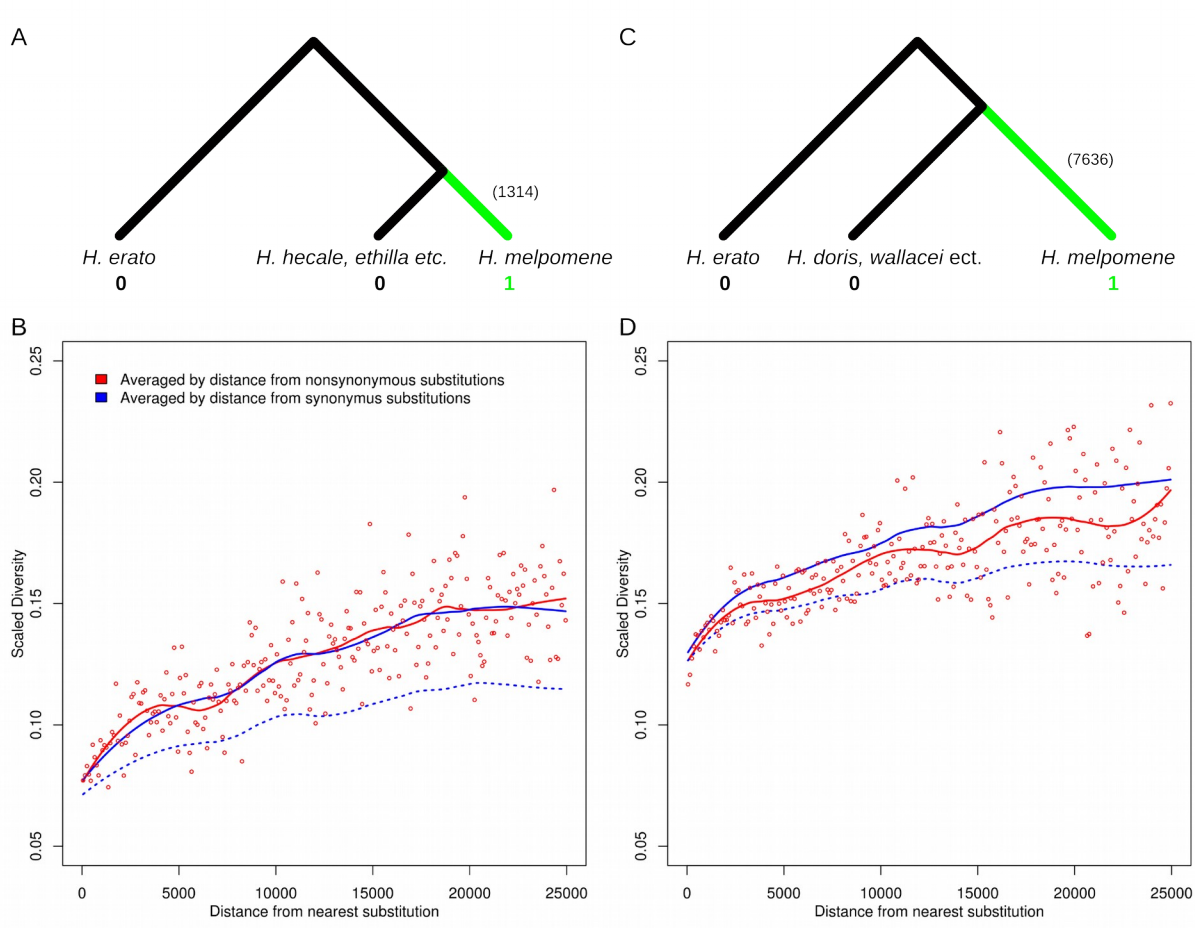

Fig. S18. Scaled diversity by distance from non-synonymous and synonymous substitutions

A, **C.** Trees showing the genotype patterns used to infer substitutions on the branch leading to *H. melpomene*, either since the split from *H. hecale* etc. (A) or the split from *H. doris* etc. (C). The number of substitutions identified is shown. **B, D.** Scaled diversity (π/d_{XY}) for four-fold degenerate (4D) sites, binned in 100 bp bins according to their distance from the nearest substitution. Red points indicate values for each bin when sites were binned by distance from non-synonymous substitutions, with the moving average (loess, span = 0.5) indicated by the red line. The solid blue line indicates the moving average when sites were binned according to their distance from synonymous substitutions and averaged over 100 bootstraps. The dashed blue line indicates the 5% quantile from the 100 bootstraps. There is no detectable reduction in scaled diversity around non-synonymous substitutions that occurred over the shorter period (B). There appears to be a slight reduction around substitutions that occurred over the longer period (D), but neither are significantly below the 95% confidence threshold.

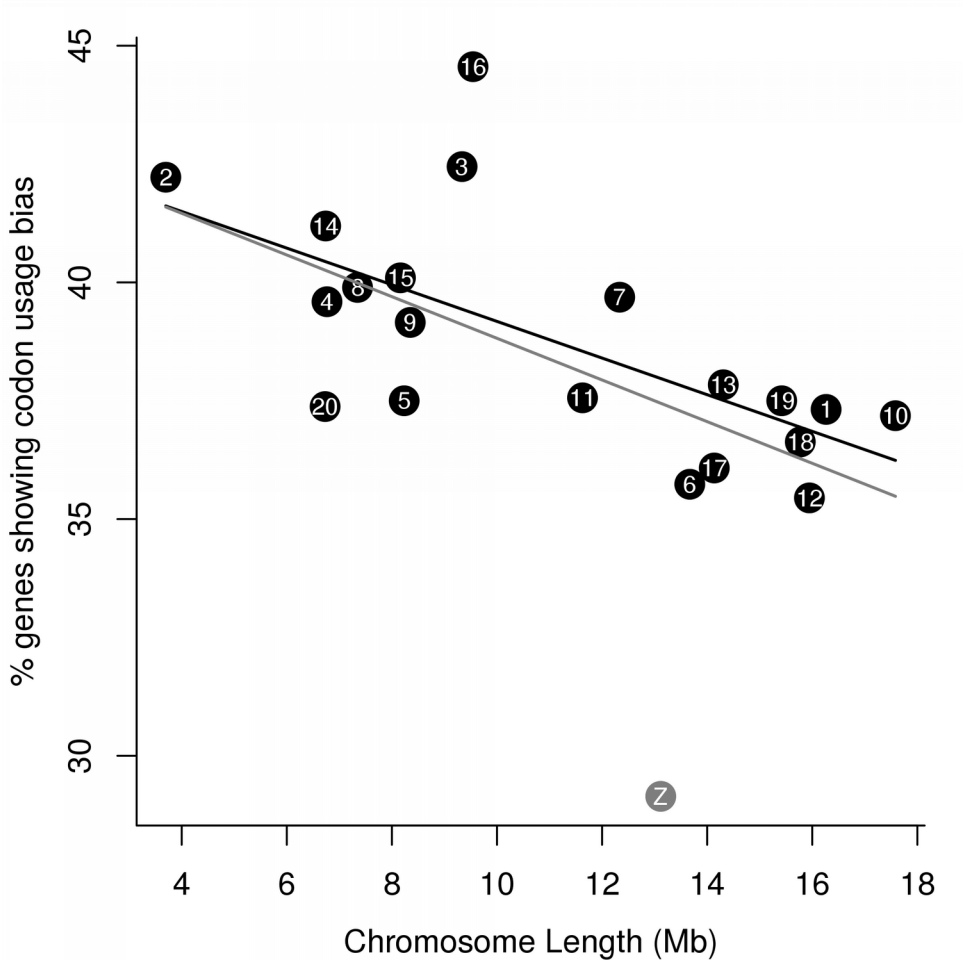

Fig. S19. Relationship between codon usage bias and chromosome length

The percentage of genes on each chromosome showing evidence of non-trivial codon usage bias (CUB), plotted against the estimated chromosome length. Chromosomes are labelled and the Z chromosome is indicated in gray. Linear regression lines are shown for all chromosomes (grey) or autosomes only (black).

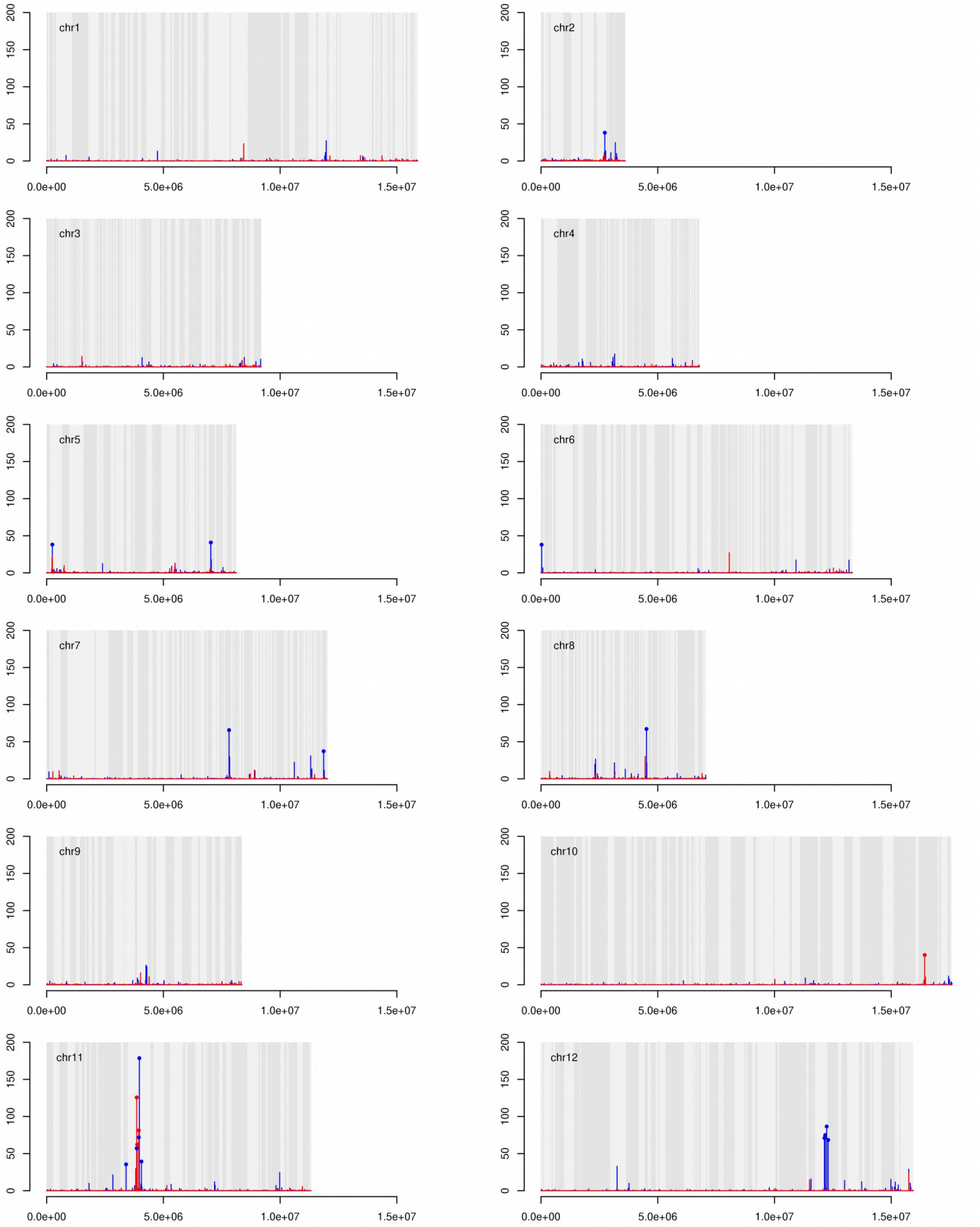

Fig. S20. CLR values from SweeD (A)

Chromosomes are shown with scaffolds shaded light and dark. Vertical lines indicate CLR values for the Western (red) and Eastern (blue) populations. Points indicate values above the cut-off value of 34, defined by analysis of simulated data.

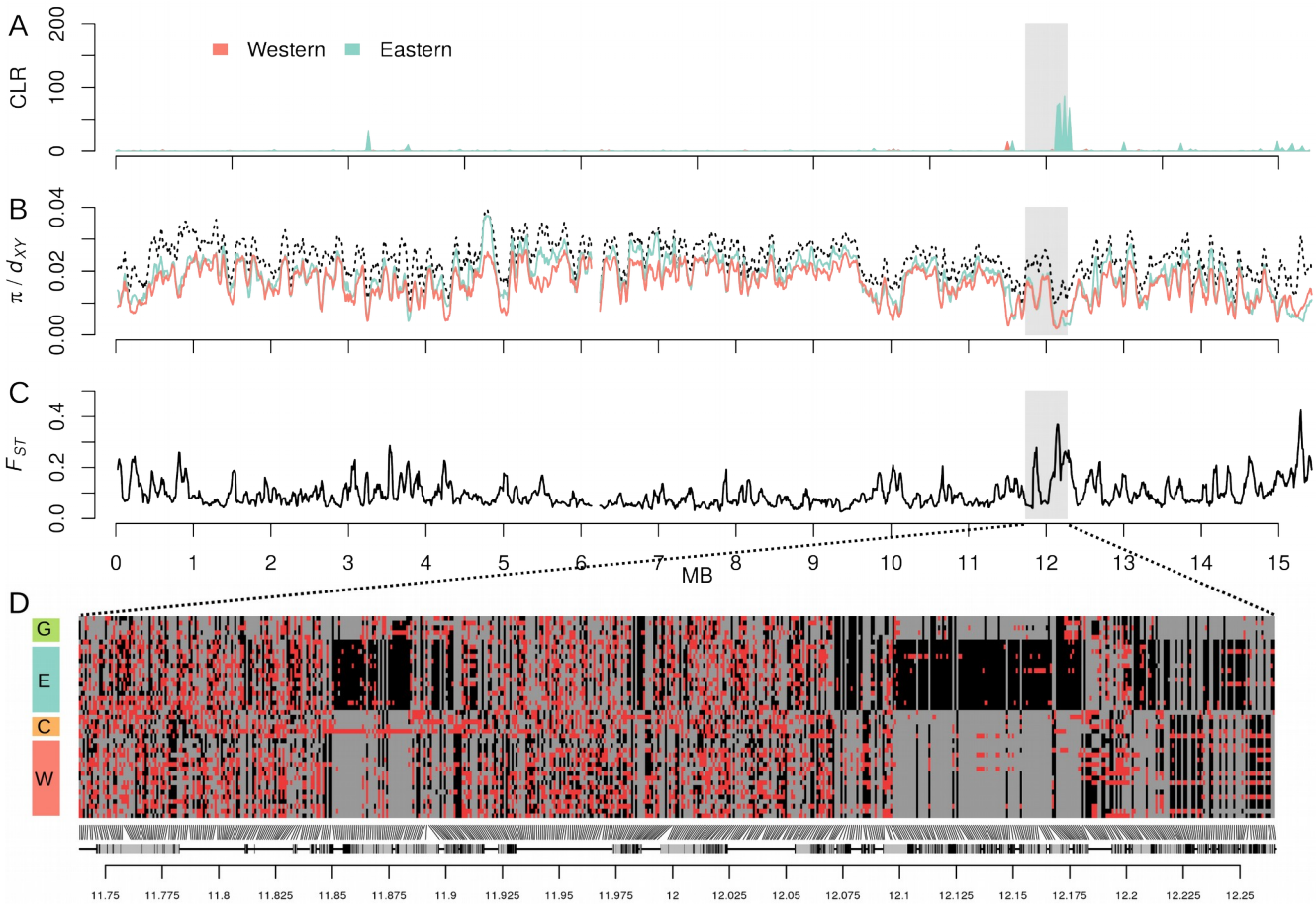

Fig. S21. A putative selective sweep on Chromosome 12

A. Composite likelihood ratio (CLR) values calculated by SweeD (Pavlidis, 2014) for the Eastern and Western populations, for 1000 windows across chromosome 12. Scaffolds are shaded light and dark. B. Nucleotide diversity (π) for the Eastern and Western populations (in colour) and divergence (d_{XY}) between these two populations (black dashed line), calculated for 50 kb sliding windows across chromosome 11, sliding in increments of 10 kb. C. F_{ST} between the Eastern and Western populations, calculated in windows as in B. D. Individual genotypes at 600 biallelic SNPs on scaffold HE671629, which harbours the putative selective sweep. Homozygous genotypes are coloured grey (major allele) and black (minor allele), and heterozygoyes are coloured red. To optimize the detection of differences between populations, SNPs with a high degree of polymorphism (minor allele frequency ≥ 0.25) were considered. The 600 SNPs plotted were sampled semi-randomly, ensuring that no two sampled SNPs were more than 1000 bp apart. Protein coding genes are indicated below the plot, with exons shown in black.

Table S1. Sample information and genotyping summary statistics

Sample ID, species and race, sex and sampling locations are given, along with genotyping summary statistics. 'Hom Ref,' 'Het' and 'Hom' Alt refer to sites that were genotyped as homozygous for the reference genome allele, heterozygous and homozygous for an alternate allele, respectively. 'ts/tv' is the ratio of transitions to transversions for non-reference (alternate) genotypes. Sequence Read Archive accession numbers are given, along with the reference for the previous study that generated the sequence data, where relevant.

Sample ID	Species and race	Sex	Lat.	Long.	population	depth	Genotyped (Mb)	% Hom Ref	% Het	% Hom Alt	ts/tv	Accession
melP.HGC1	*H. m. melpomene*		*NA*	NA	*NA*	17.95	*254.5*	99.39	*0.61*	0	*1.25*	SRR424576
ros.MK523	*H. m. rosina*	M	*9.717*	-83.05	*West*	17.22	*218.9*	97.7	*1.49*	0.82	*1.27*	SRS518836
ros.MK524	*H. m. rosina*	F	*9.85*	-84.317	*West*	17.24	*217.6*	97.63	*1.62*	0.75	*1.27*	SRS518837
ros.MK525	*H. m. rosina*	M	*8.467*	-83.583	*West*	16.87	*219.2*	97.64	*1.64*	0.72	*1.27*	SRS518838
ros.MK589	*H. m. rosina*	M	*9.867*	-83	*West*	17.07	*216.8*	97.78	*1.45*	0.77	*1.25*	SRS518839
ros.MK675	*H. m. rosina*	M	*9.4*	-84.167	*West*	17.40	*219.7*	97.6	*1.63*	0.77	*1.26*	SRS518840
ros.MK676	*H. m. rosina*	M	*9.717*	-83.05	*West*	15.87	*214.4*	97.85	*1.4*	0.75	*1.27*	SRS518841
ros.MK682	*H. m. rosina*	M	*10.433*	-83.983	*West*	17.34	*218.2*	97.74	*1.41*	0.85	*1.27*	SRS519003
ros.MK683	*H. m. rosina*	M	*9.4*	-84.167	*West*	16.58	*217.2*	97.69	*1.59*	0.72	*1.26*	SRS518842
ros.MK687	*H. m. rosina*	F	*10.433*	-83.983	*West*	17.17	*218.2*	97.74	*1.4*	0.86	*1.26*	SRS518843
ros.MK689	*H. m. rosina*	F	*9.85*	-84.317	*West*	16.86	*217.5*	97.66	*1.6*	0.74	*1.27*	SRS518844
ros.CJ531	*H. m. rosina*	M	*9.121*	-79.697	*West*	29.24	*208.9*	97.54	*1.73*	0.73	*1.32*	ERR260277
ros.CJ533	*H. m. rosina*	M	*9.121*	-79.697	*West*	29.81	*203.1*	97.56	*1.74*	0.7	*1.33*	ERR260278
ros.CJ546	*H. m. rosina*	M	*9.121*	-79.697	*West*	29.42	*203.7*	97.58	*1.74*	0.69	*1.33*	ERR260279
ros.CJ2071	*H. m. rosina*	M	*9.121*	-79.697	*West*	37.69	*224.9*	97.32	*1.82*	0.86	*1.28*	ERR260280
melP.CJ18038	*H. m. melpomene*	F	*8.614*	-78.14	*West*	58.80	*228.2*	97.17	*1.88*	0.95	*1.28*	ERR260285
melP.CJ18097	*H. m. melpomene*	M	*8.28*	-77.81	*West*	17.58	*188.7*	97.76	*1.7*	0.55	*1.35*	ERR260286
vul.CJ14632	*H. m. vulcanus*	M	*8.614*	-78.14	*West*	19.18	*125*	97.91	*1.57*	0.52	*1.42*	ERS977709

vul.CS519	*H. m. vulcanus*	M	*3.9*	-76.633	*West*	15.95	*188.5*	97.87	*1.55*	0.58	*1.34*	ERS977710
vul.CS10	*H. m. vulcanus*	M	*3.9*	-76.633	*West*	36.63	*231.4*	97.21	*1.83*	0.96	*1.25*	ERS1030540
vul.CS11	*H. m. vulcanus*	M	*3.9*	-76.633	*West*	9.95	*192.9*	98.28	*1.49*	0.23	*1.25*	ERS1030541
cyth.CJ2856	*H. m. cythera*	M	*-0.32*	-79.337	*West*	27.08	*202.3*	97.35	*1.91*	0.74	*1.33*	ERS977691
cyth.CJ2857	*H. m. cythera*	M	*-0.32*	-79.337	*West*	18.97	*92.4*	98.18	*1.39*	0.43	*1.47*	ERS977692
melC.CS3	*H. m. melpomene*	M	*4.213*	-73.803	*Colombia*	43.04	*229.2*	96.96	*1.91*	1.13	*1.26*	ERS1030542
melC.CS6	*H. m. melpomene*	M	*5.617*	-72.3	*Colombia*	106.39	*237.2*	96.75	*2.14*	1.11	*1.24*	ERS1030543
melC.CS25	*H. m. melpomene*	M	*4.213*	-73.803	*Colombia*	26.95	*223.4*	97.04	*1.87*	1.09	*1.26*	ERS1030549
melC.CS26	*H. m. melpomene*	M	*4.213*	-73.803	*Colombia*	25.98	*222.9*	97.11	*1.85*	1.04	*1.27*	ERS1030550
melC.CS27	*H. m. melpomene*	M	*5.617*	-72.3	*Colombia*	42.59	*227.9*	96.88	*2.04*	1.09	*1.25*	ERS1030551
moc.CS228	*H. m. mocoa*	M	*1.178*	-76.665	*East*	21.19	*195.2*	97.07	*1.84*	1.09	*1.34*	ERS977689
moc.CS231	*H. m. mocoa*	M	*1.178*	-76.665	*East*	9.24	*95.6*	98.4	*1.37*	0.22	*1.43*	ERS977694
moc.CS16	*H. m. mocoa*	M	*1.178*	-76.665	*East*	24.22	*214.1*	96.67	*1.97*	1.36	*1.27*	ERS1030544
moc.CS17	*H. m. mocoa*	M	*1.178*	-76.665	*East*	31.53	*216.6*	96.62	*2.14*	1.23	*1.26*	ERS1030545
ple.CJ9156	*H. m. plesseni*	F	*-1.398*	-78.178	*East*	18.99	*192.2*	97.24	*1.69*	1.07	*1.33*	ERS977705
ple.CJ16293	*H. m. plesseni*	F	*-1.46*	-78.073	*East*	19.87	*117.6*	97.62	*1.58*	0.81	*1.41*	ERS977706
mapl.CJ16042	*H. m. plesseniX malleti*	M	*-1.371*	-77.875	*East*	22.71	*195.7*	97.02	*1.87*	1.11	*1.34*	ERS977697
mal.CJ16550	*H. m. malleti*	M	*-1.061*	-77.668	*East*	18.22	*105*	97.61	*1.63*	0.76	*1.41*	ERS977698
mal.CJ17162	*H. m. malleti*	F	*-1.061*	-77.668	*East*	30.13	*184.9*	97.02	*1.83*	1.15	*1.36*	ERS977699
mal.CS21	*H. m. malleti*	M	*1.814*	-75.669	*East*	34.16	*218.2*	96.4	*2.09*	1.51	*1.26*	ERS1030546
mal.CS22	*H. m. malleti*	M	*1.61*	-75.667	*East*	30.92	*217.2*	96.41	*2.1*	1.49	*1.26*	ERS1030547
mal.CS24	*H. m. malleti*	M	*1.751*	-75.632	*East*	26.15	*215.4*	96.51	*2.06*	1.43	*1.27*	ERS1030548
ecu.CJ9117	*H. m. ecuadorensis*	M	*-4.059*	-78.955	*East*	22.56	*187.8*	97.07	*1.81*	1.12	*1.34*	ERS977695

ecu.CJ9121	*H. m. ecuadorensis*	M	*-3.059*	-77.955	*East*	23.40	*137.1*	97.39	*1.65*	0.96	*1.42*	ERS977696
ama.JM216	*H. m. amaryllis*	M	*-5.676*	-77.675	*East*	31.00	*218.1*	96.44	*2.03*	1.53	*1.27*	ERR260289
ama.JM160	*H. m. amaryllis*	F	*-6.469*	-76.353	*East*	42.10	*221.3*	96.31	*2.02*	1.66	*1.26*	ERR260289
ama.JM293	*H. m. amaryllis*	F	*-6.47*	-76.347	*East*	50.92	*223.1*	96.25	*2.05*	1.7	*1.25*	ERR260290
ama.JM48	*H. m. amaryllis*	M	*-6.096*	-76.977	*East*	53.05	*223.7*	96.19	*2.14*	1.67	*1.25*	ERR260287
agl.JM108	*H. m. aglaope*	M	*-5.91*	-76.226	*East*	34.34	*218*	96.48	*2.02*	1.51	*1.28*	ERR260291
agl.JM112	*H. m. aglaope*	M	*-5.91*	-76.226	*East*	37.23	*220.3*	96.34	*2.07*	1.59	*1.26*	ERR260292
agl.JM569	*H. m. aglaope*	M	*-5.946*	-76.245	*East*	42.32	*221.5*	96.31	*2.08*	1.61	*1.26*	ERR260293
agl.JM572	*H. m. aglaope*	M	*-5.946*	-76.247	*East*	35.56	*219.4*	96.39	*2.06*	1.55	*1.27*	ERR260294
aman.CS2228	*H. m. amandus*	M	*-18.158*	-63.509	*East*	20.03	*204.4*	96.96	*1.9*	1.14	*1.32*	ERS977689
aman.CS2221	*H. m. amandus*	M	*-18.158*	-63.509	*East*	50.79	*201.4*	96.62	*1.9*	1.49	*1.32*	ERS977688
melG.CJ9315	*H. m. melpomene*	M	*4.963*	-52.42	*Guiana*	25.00	*204.3*	97.15	*1.56*	1.29	*1.34*	ERR260281
melG.CJ9316	*H. m. melpomene*	M	*4.963*	-52.42	*Guiana*	24.27	*199.6*	97.26	*1.55*	1.18	*1.35*	ERR260282
melG.CJ9317	*H. m. melpomene*	M	*4.963*	-52.42	*Guiana*	35.24	*213.3*	96.8	*1.63*	1.58	*1.31*	ERR260283
melG.CJ13435	*H. m. melpomene*	M	*4.915*	-52.376	*Guiana*	36.14	*213.7*	96.76	*1.66*	1.58	*1.31*	ERR260284
thel.CJ13566	*H. m. thelxiopeia*	M	*3.656*	-54.039	*Guiana*	22.97	*204.9*	97.08	*1.54*	1.38	*1.33*	ERS977708
mer.CJ13715	*H. m. meriana*	M	*3.656*	-54.039	*Guiana*	11.12	*65*	98.41	*1.26*	0.33	*1.39*	ERS977704
mer.CJ13819	*H. m. meriana*	M	*3.656*	-54.039	*Guiana*	21.30	*145.8*	97.53	*1.41*	1.06	*1.42*	ERS977703
cyd.CJ553	*H. cydno chioneus*	M	*9.171*	-79.757	*NA*	37.18	*211*	96.43	*2.03*	1.54	*1.27*	ERR260295
cyd.CJ560	*H. cydno chioneus*	M	*9.171*	-79.757	*NA*	36.59	*211.4*	96.45	*1.98*	1.57	*1.27*	ERR260296
cyd.CJ564	*H. cydno*	M	*9.171*	-79.757	*NA*	40.50	*211.3*	96.4	*2.06*	1.54	*1.27*	ERR260297

	chioneus											
cyd.CJ565	*H. cydno chioneus*	M	*9.171*	-79.757	*NA*	47.41	*214.8*	96.25	*2.1*	1.65	*1.25*	ERR260298
tim.JM313	*H. timareta thelxinoe*	M	*-6.453*	-76.289	*NA*	39.73	*219.6*	96.46	*1.63*	1.92	*1.25*	ERR260302
tim.JM57	*H. timareta thelxinoe*	M	*-6.455*	-76.298	*NA*	42.86	*220*	96.44	*1.66*	1.9	*1.25*	ERR260299
tim.JM84	*H. timareta thelxinoe*	M	*-6.455*	-76.298	*NA*	30.23	*215.5*	96.64	*1.61*	1.74	*1.27*	ERR260300
tim.JM86	*H. timareta thelxinoe*	M	*-6.455*	-76.298	*NA*	42.86	*220.6*	96.41	*1.69*	1.91	*1.25*	ERR260301
hec.JM273	*H. hecale felix*	F	*-5.972*	-76.232	*NA*	38.98	*206.6*	95.56	*1.67*	2.77	*1.26*	ERR260306
eth.JM67	*H. ethilla aerotome*	M	*-6.467*	-76.335	*NA*	43.70	*204.9*	95.7	*0.89*	3.41	*1.25*	ERR260305
par.JM371	*H. pardalinus ssp. nov.*	M	*-8.343*	-74.592	*NA*	37.86	*205.7*	95.02	*2.85*	2.12	*1.27*	ERR260303
ser.JM202	*H. pardalinus sergestus*	M	*-6.478*	-76.352	*NA*	37.53	*205*	96.05	*1.2*	2.75	*1.26*	ERR260304
wal.CJ8687	*H. wallacei*	M	*-6.29*	-76.229	*NA*	14.06	*88.4*	96.59	*1.19*	2.22	*1.44*	ERS1030552
bur.CJ8560	*H. burneyi*	M	*-6.458*	-76.288	*NA*	19.32	*73.6*	96.41	*1.12*	2.48	*1.48*	ERS1030553
hecu.CJ8550	*H. hecuba*	M	*0.475*	-77.555	*NA*	16.99	*90*	96.88	*0.41*	2.71	*1.4*	ERS977683
hier.CJ8149	*H. hierax*	M	*-0.182*	-77.685	*NA*	10.60	*65.4*	98.62	*0.26*	1.12	*1.51*	ERS977685
dor.JC8684	*H. doris*	M	*-6.29*	-76.2289	*NA*	16.29	*83*	96.59	*0.56*	2.84	*1.42*	ERS977668
era.CJ2979	*H. erato*	M	*9.145*	-79.729	*NA*	36.96	*108.3*	94.09	*1.83*	4.08	*1.43*	ERS1030555
era.CJ2980	*H. erato*	M	*9.145*	-79.729	*NA*	35.85	*107*	94.17	*1.83*	3.99	*1.44*	ERS1030556
era.CJ2981	*H. erato*	M	*9.145*	-79.729	*NA*	29.47	*101.4*	94.74	*1.73*	3.54	*1.48*	ERS1030557
era.CJ618	*H. erato*	M	*9.122*	-79.715	*NA*	33.79	*105.8*	94.36	*1.79*	3.85	*1.45*	ERS1030554
						29.5	***189.3***	**96.92**	***1.68***	**1.40**	***1.31***	

Table S2. Comparing alternative genotyping approaches

Summary statistics and estimated error rates for different genotyping approaches: with or without indel realignment or simultaneous genotyoing. '≥40/80 called' refers to the number of sites with high-quality genotype calls for at least 40 of the 80 samples.

Indel realignment	Called simultaneously	Average calls per sample (%)	≥40/80 called	≥60/80 called	80/80 called	ts/tv	Est. error rate
No	No	68.7	79.4	54.9	1.2	1.38	0.00032
No	Yes	35.2	40.6	28.9	8.4	1.40	0.00084
Yes	No	68.7	79.4	54.9	1.2	1.38	0.00032
Yes	Yes	35.2	40.6	28.9	8.4	1.40	0.00084

Table S3. Nucleotide diversity (π) and absolute divergence (d_{XY}), means with errors

An extended version of Table 1, including the following site classes: All sites, intergenic, intronic, codon positions 1, 2 and 3, and 4D sites. Standard deviations and standard errors are given for all values. CV refers to the coefficient of variation, the ratio of the mean to the standard deviation. 'WG' values are for the whole genome and 'Z' values are specifically for the Z chromosome.

site class	region	measure	mean	sd	se	CV
All	WG	pi (*melpomene*)	0.0192	0.00631	0.00010	0.329
All	WG	dXY (*melpomene* – *cydno* clade)	0.0275	0.00532	0.00008	0.194
All	WG	dXY (*melpomene* – *silvaniform* clade)	0.0361	0.00604	0.00010	0.167
All	Z	pi (*melpomene*)	0.0109	0.00462	0.00036	0.425
All	Z	dXY (*melpomene* – *cydno* clade)	0.0241	0.00447	0.00034	0.185
All	Z	dXY (*melpomene* – *silvaniform* clade)	0.0337	0.00570	0.00044	0.169
Intergenic	WG	pi (*melpomene*)	0.0204	0.00631	0.00010	0.310
Intergenic	WG	dXY (*melpomene* – *cydno* clade)	0.0288	0.00525	0.00008	0.182
Intergenic	WG	dXY (*melpomene* – *silvaniform* clade)	0.0377	0.00619	0.00010	0.164
Intergenic	Z	pi (*melpomene*)	0.0118	0.00482	0.00038	0.410
Intergenic	Z	dXY (*melpomene* – *cydno* clade)	0.0253	0.00445	0.00035	0.176
Intergenic	Z	dXY (*melpomene* – *silvaniform* clade)	0.0353	0.00573	0.00045	0.162
Intron	WG	pi (*melpomene*)	0.0185	0.00649	0.00013	0.351
Intron	WG	dXY (*melpomene* – *cydno* clade)	0.0285	0.00569	0.00011	0.200
Intron	WG	dXY (*melpomene* – *silvaniform* clade)	0.0382	0.00597	0.00012	0.156
Intron	Z	pi (*melpomene*)	0.0107	0.00395	0.00034	0.371
Intron	Z	dXY (*melpomene* – *cydno* clade)	0.0254	0.00444	0.00038	0.175
Intron	Z	dXY (*melpomene* – *silvaniform* clade)	0.0355	0.00476	0.00041	0.134

Codon 1	WG	pi (*melpomene*)	0.0064	0.00521	0.00011	0.812
Codon 1	WG	dXY (*melpomene* – *cydno* clade)	0.0102	0.00664	0.00014	0.649
Codon 1	WG	dXY (*melpomene* – *silvaniform* clade)	0.0137	0.00772	0.00016	0.563
Codon 1	WG	dXY (*melpomene* – *wallacei* clade)	0.0255	0.01448	0.00030	0.567
Codon 1	WG	dXY (*melpomene* – *doris* clade)	0.0224	0.01251	0.00026	0.559
Codon 1	WG	dXY (*melpomene* – *erato*)	0.0366	0.01823	0.00038	0.498
Codon 1	Z	pi (*melpomene*)	0.0029	0.00168	0.00016	0.573
Codon 1	Z	dXY (*melpomene* – *cydno* clade)	0.0083	0.00433	0.00040	0.522
Codon 1	Z	dXY (*melpomene* – *silvaniform* clade)	0.0135	0.00698	0.00065	0.518
Codon 1	Z	dXY (*melpomene* – *wallacei* clade)	0.0280	0.01433	0.00133	0.513
Codon 1	Z	dXY (*melpomene* – *doris* clade)	0.0238	0.01242	0.00115	0.521
Codon 1	Z	dXY (*melpomene* – *erato*)	0.0362	0.01734	0.00161	0.480
Codon 2	WG	pi (*melpomene*)	0.0057	0.00478	0.00010	0.833
Codon 2	WG	dXY (*melpomene* – *cydno* clade)	0.0091	0.00645	0.00013	0.707
Codon 2	WG	dXY (*melpomene* – *silvaniform* clade)	0.0122	0.00794	0.00016	0.650
Codon 2	WG	dXY (*melpomene* – *wallacei* clade)	0.0223	0.01581	0.00033	0.709
Codon 2	WG	dXY (*melpomene* – *doris* clade)	0.0198	0.01379	0.00029	0.695
Codon 2	WG	dXY (*melpomene* – *erato*)	0.0323	0.02019	0.00042	0.625
Codon 2	Z	pi (*melpomene*)	0.0025	0.00169	0.00016	0.667
Codon 2	Z	dXY (*melpomene* – *cydno* clade)	0.0072	0.00513	0.00048	0.710
Codon 2	Z	dXY (*melpomene* – *silvaniform* clade)	0.0116	0.00697	0.00065	0.599
Codon 2	Z	dXY (*melpomene* – *wallacei* clade)	0.0226	0.01411	0.00131	0.625
Codon 2	Z	dXY (*melpomene* – *doris* clade)	0.0206	0.01344	0.00125	0.652
Codon 2	Z	dXY (*melpomene* – *erato*)	0.0300	0.01823	0.00169	0.607
Codon 3	WG	pi (*melpomene*)	0.0150	0.00924	0.00019	0.615
Codon 3	WG	dXY (*melpomene* – *cydno* clade)	0.0239	0.01096	0.00023	0.458
Codon 3	WG	dXY (*melpomene* – *silvaniform* clade)	0.0326	0.01263	0.00026	0.387
Codon 3	WG	dXY (*melpomene* – *wallacei* clade)	0.0645	0.02446	0.00051	0.379
Codon 3	WG	dXY (*melpomene* – *doris* clade)	0.0564	0.02144	0.00044	0.380
Codon 3	WG	dXY (*melpomene* – *erato*)	0.0909	0.02881	0.00060	0.317
Codon 3	Z	pi (*melpomene*)	0.0071	0.00379	0.00035	0.536
Codon 3	Z	dXY (*melpomene* – *cydno* clade)	0.0194	0.00749	0.00070	0.385
Codon 3	Z	dXY (*melpomene* – *silvaniform* clade)	0.0304	0.00978	0.00091	0.321
Codon 3	Z	dXY (*melpomene* – *wallacei* clade)	0.0643	0.02204	0.00205	0.343
Codon 3	Z	dXY (*melpomene* – doris clade)	0.0586	0.01987	0.00185	0.339
Codon 3	Z	dXY (*melpomene* – *erato*)	0.0937	0.02988	0.00277	0.319

4D	WG	pi (*melpomene*)	0.0251	0.01332	0.00029	0.531
4D	WG	dXY (*melpomene* – *cydno* clade)	0.0409	0.01472	0.00032	0.360
4D	WG	dXY (*melpomene* – *silvaniform* clade)	0.0565	0.01605	0.00035	0.284
4D	WG	dXY (*melpomene* – *wallacei* clade)	0.1136	0.02369	0.00052	0.208
4D	WG	dXY (*melpomene* – *doris* clade)	0.0996	0.02111	0.00046	0.212
4D	WG	dXY (*melpomene* – *erato*)	0.1579	0.02894	0.00063	0.183
4D	Z	pi (*melpomene*)	0.0112	0.00479	0.00045	0.427
4D	Z	dXY (*melpomene* – *cydno* clade)	0.0315	0.00886	0.00084	0.281
4D	Z	dXY (*melpomene* – *silvaniform* clade)	0.0493	0.01141	0.00108	0.231
4D	Z	dXY (*melpomene* – *wallacei* clade)	0.1073	0.01923	0.00182	0.179
4D	Z	dXY (*melpomene* – *doris* clade)	0.1003	0.02103	0.00200	0.210
4D	Z	dXY (*melpomene* – *erato*)	0.1578	0.02964	0.00281	0.188

Table S4. Correlation Coefficients between all variables in the multiple regression main model

	pi	d_S	D_n	gene density	GC content	$\hat{r}$
pi	1	0.39	-0.18	-0.39	-0.09	0.19
d_S	0.39	1	0.23	-0.11	-0.14	-0.01
D_n	-0.18	0.23	1	0.53	-0.23	-0.1
gene density	-0.39	-0.11	0.53	1	-0.13	-0.08
GC content	-0.09	-0.14	-0.23	-0.13	1	0
$\hat{r}$	0.19	-0.01	-0.1	-0.08	0	1

Table S5. Summary of multiple regression with five explanatory variables for 4D site diversity, calculated for 100 kb windows, excluding *H. erato* ($R^2 = 0.267$; adjusted $R^2 = 0.264$; $F_{5,1461} = 106.2$; $p < 2.2e\text{-}16$).

	Estimate	Std. error	SS	RSS	$F_{1, 1461}$	P (>F)	Partial R^2	VIF
Gene den.	-0.0137	0.0010	146.451	1379.4	173.544	**< 2.2e-16**	0.1062	1.456
$\hat{r}$	0.0062	0.0008	49.506	1282.4	58.664	**3.384e-14**	0.0386	1.009
D_n	-0.0049	0.0011	18.573	1251.5	22.009	**2.968e-06**	0.0148	1.679
d_S	0.0090	0.0010	73.843	1306.8	87.504	**< 2.2e-16**	0.0565	1.212
GC cont.	-0.0050	0.0009	28.055	1261.0	33.245	**9.896e-09**	0.0222	1.063
(Intercept)	0.1520	0.0009						

$\hat{r}$: recombination rate, D_n: number of non-synonymous substitutions, d_S: synonymous substitutions per synonymous site, SS: Sum of Squares, RSS: Residual Sum of Squares, VIF: variance inflation factor

Table S6. Summary of multiple regression with five explanatory variables for 4D site diversity, calculated for 100 kb windows using *a* after removal of the upper and lower 10th percentile for *a* (R^2 = 0.321; adjusted R^2 = 0.318; $F_{5,1167}$ = 110.3; p < 2.2e-16).

	Estimate	Std. error	SS	RSS	$F_{1, 1167}$	P (>F)	Partial R^2	VIF
Gene den.	-0.0102	0.0009	79.022	859.03	118.228	**< 2.2e-16**	0.0920	1.111
$\hat{r}$	0.0059	0.0008	33.701	813.71	50.421	**2.151e-12**	0.0414	1.005
a	-0.0072	0.0008	49.598	829.61	74.206	**< 2.2e-16**	0.0598	1.086
d_S	0.0120	0.00o9	108.931	888.94	162.976	**< 2.2e-16**	0.1225	1.040
GC cont.	-0.0032	0.0009	8.689	788.70	12.999	**0.00032**	0.0110	1.048
(Intercept)	0.1531	0.0009						

a: number of adaptive substitutions

Table S7. Summary of multiple regression with five explanatory variables for 4D site diversity, calculated for 100 kb windows, including only genes with minimal codon usage bias (R^2 = 0. 0.327; adjusted R^2 = 0.325; $F_{5,1393}$ = 135.5; p < 2.2e-16).

	Estimate	Std. error	SS	RSS	$F_{1, 1393}$	P (>F)	Partial R^2	VIF
Gene den.	-0.0132	0.0011	86.485	928.40	143.095	**< 2.2e-16**	0.0932	1.521
$\hat{r}$	0.0048	0.0008	20.653	862.57	34.171	**6.278e-09**	0.0239	1.011
D_n	-0.0042	0.0010	10.274	852.19	16.999	**3.960e-05**	0.0121	1.585
d_S	0.0154	0.0010	148.035	989.95	244.933	**< 2.2e-16**	0.1495	1.153
GC cont.	-0.0039	0.0009	0.166	842.08	0.2742	0.601	0.0002	1.076
(Intercept)	0.1523	0.0009						

Table S8. Summary of multiple regression with five explanatory variables for 4D site diversity, calculated for 100 kb windows, excluding chromosome ends, cut off = 0.9 (R^2 = 0.350; adjusted R^2 = 0.347; $F_{5,1330}$ = 143; p < 2.2e-16).

	Estimate	Std. error	SS	RSS	$F_{1, 1330}$	P (>F)	Partial R^2	VIF
Gene den.	-0.0147	0.0010	147.410	1105.99	204.527	**< 2.2e-16**	0.1333	1.503
$\hat{f}$	0.0050	0.0008	29.297	987.88	40.649	**2.507e-10**	0.0297	1.014
D_n	-0.0024	0.0010	4.063	962.64	5.637	**0.01773**	0.0042	1.587
d_S	0.0132	0.0009	144.512	1103.09	200.506	**< 2.2e-16**	0.1310	1.167
GC cont.	-0.0010	0.0008	1.156	959.74	1.604	0.20556	0.0012	1.066
(Intercept)	0.1546	0.0009						

Table S9. BLAST hits for genes in putative selective sweep regions

For each gene in the two putative sweep intervals, the start and end positions on the Hmel1.1 scaffold and chromosome are given, along with details of the best blast hit in *D. melanogaster*, where applicable.

Hmel1-1 Gene	Start on scaffold	End on scaffold	Start on Chrom	End on Chrom	E-value	hit	Predictions/Assertions
Chromosome 11 – HE672079							
HMEL015151	*320191*	323579	*3786671*	3790059	*7.74E-053*	CG18003	FMN binding;glycolate oxidase activity
HMEL015153	*328330*	334579	*3794810*	3801059	*0*	CG11964	
HMEL015154/5/6	*334035*	355595	*3800515*	3822075	*0*	Gclc	glutamate-cysteine ligase activity
HMEL015157	*356497*	374505	*3808419*	3813015	*9.23E-144*	MED14	RNA polymerase II transcription cofactor activity
HMEL015158	*374565*	375639	*3841045*	3842119	*5.39E-007*	CG15279	cation:amino acid symporter activity; neurotransmitter transporter activity; neurotransmitter:sodium symporter activity
HMEL015159	*380092*	396226	*3846572*	3862706	*4.16E-166*	pbl	GTPase activator activity; guanyl-nucleotide exchange factor activity; Rho guanyl-nucleotide exchange factor activity
HMEL015160	*403996*	431878	*3870476*	3898358	*3.69E-144*	Ugt	UDP-glucose:glycoprotein glucosyltransferase activity
HMEL015161	*404077*	408880	*3870557*	3875360	*4.66E-040*	Tango11	
HMEL015162	*433587*	435190	*3900067*	3901670	*0.000329717*	Wnt5	frizzled-2 binding; receptor binding
HMEL015163	*451928*	463252	*3918408*	3929732	*4.17E-123*	Taf6	protein heterodimerization activity
HMEL015164	*464608*	477699	*3931088*	3944179	*4.33E-167*	Nup358	Ran GTPase binding; zinc ion binding
HMEL015165	*480092*	483190	*3946572*	3949670	*7.81E-008*	CG8765	
HMEL015166	*491653*	496970	*3958133*	3963450	*1.01E-054*	SelR	peptide-methionine (R)-S-oxide reductase activity; zinc ion binding
HMEL015167	*491666*	504840	*3958146*	3971320	*1.34E-174*	Pak	ATP binding; protein serine/threonine kinase activity; receptor signaling protein serine/threonine kinase activity
HMEL015168	*505801*	519256	*3972281*	3985736	*6.24E-054*	GCC185	identical protein binding; protein homodimerization activity
HMEL015169	*520810*	532469	*3987290*	3998949	*0.00E+000*	Ef1alpha 100E	GTP binding; GTPase activity; translation elongation factor activity

Chromosome 12 – HE671629

HMEL017562	*85336*	92585	*12183400*	12176151	*4.37E-163*	CG4678	metallocarboxypeptidase activity; serine-type carboxypeptidase activity; zinc ion binding
HMEL017563	*89216*	89806	*12179520*	12178930	*5.53E-033*	Sod3	copper ion binding; superoxide dismutase activity; zinc ion binding
HMEL017564	*94058*	102910	*12174678*	12165826	*8.08E-059*	RfC3	ATP binding; DNA binding; nucleoside-triphosphatase activity
HMEL017565	*103878*	115895	*12164858*	12152841	*3.03E-071*	CG6379	methyltransferase activity; nucleic acid binding
HMEL017566	*117818*	119977	*12150918*	12148759	*1.14E-047*	Ref1	mRNA binding; nucleic acid binding; nucleotide binding; transcription coactivator activity
HMEL017567	*120263*	124745	*12148473*	12143991	*1.58E-044*	CG7544	methyltransferase activity; nucleic acid binding
HMEL017568	*125619*	128357	*12143117*	12140379	*2.69E-021*	Slbp	mRNA binding
HMEL017569	*127478*	132204	*12141258*	12136532	*1.36E-017*	Hmg-2	DNA binding
HMEL017570	*132814*	147340	*12135922*	12121396	*2.24E-016*	ebo	Ran GTPase binding
HMEL017572	*143815*	160137	*12124921*	12108599	*2.52E-164*	Cap-D2	
HMEL017573	*161080*	165195	*12107656*	12103541	*3.07E-070*	CG5885	
HMEL017574	*162409*	176005	*12106327*	12092731	*2.03E-086*	CG32447	G-protein coupled receptor activity; glutamate receptor activity
HMEL017575	*179758*	182846	*12088978*	12085890	*1.04E-048*	CG12024	

SUPPLEMENTARY METHODS

Mapping and genotyping

Reads were mapped using Stampy v1.0 (Lunter and Goodson 2011). Defaults were used for all parameters with the exception of the expected substitution rate, which depended on the species of the sample being mapped: *H. melpomene* = 0.03, *H. cydno*/*timareta* = 0.04, *H. hecale*/*ethilla*/*pardalinus* = 0.05, *H. wallacei*/*burneyi*/*doris*/*hecuba*/*hierax* = 0.07 and *H. erato* = 0.1. However, this value made little difference to the number of reads mapped and the number of SNPs genotyped (data not shown). Base alignment quality (BAQ) was considered during mapping. Local realignment around indels was performed using the Genome Analysis Tool Kit (GATK) v2.7 (DePristo *et al.* 2011). SAM/BAM file conversion, analysis and filtering were performed using SAMtools (Li *et al.* 2009) and Picard (http://picard.sourceforge.net). PCR-duplicate reads were removed using Picard.

Genotypes were called using the GATK v2.7 UnifiedGenotyper (DePristo *et al.* 2011). Default parameters were used, except for expected heterozygosity, which was set to 0.01, and BAQ calculation was performed where necessary to optimize calls around indels. We tested two different genotyping procedures: samples were either genotyped independently, or simultaneously. The former was eventually selected for all downstream analyses.

Estimation of genotyping error rates

Although we lacked Sanger-sequencing data for any of the sampled individuals, we were able to estimate the error rate in our genotyping pipeline using two separate approaches. We estimated the rate of false-positive SNP calls by examining genotypes of an individual from the inbred (five generations) *H. melpomene* reference genome strain. We found that this individual (melP.HGC1) had large tracts of homozygosity (Fig. S2). In particular the whole of chromosome 2 appears to be homozygous. We therefore took the proportion of heterozygous genotype calls on scaffolds that mapped to chromosome 2 as a proxy for the false-positive SNP discovery rate.

As the above approach is only informative about genotyping of samples very closely related to the reference, we also estimated genotyping error rates using simulated reads representing a range of divergences and sequencing depths. First, we used seq-gen (Rambaut and Grass 1997) to

simulate a 1Mb reference sequences with a GC content of 32%, along with diploid sequences (2% heterozygosity) for other taxa at increasing levels of divergence (2, 4, 6, 8 and 10%). Paired end reads of 100 bp were simulated for each taxon using ART (Huang *et al.* 2012), with an error profile mimicking that of the Illumina HiSeq 2000. For each taxon, we simulated four sets of reads, which could be combined to produce mappings of 10x, 20x, 30x and 40x. We then used our genotyping pipeline to call genotypes and compared these to the true genotype for each taxon.

Site annotation and codon usage

The *Heliconius melpomene* reference genome v1.1 annotation (The Heliconius Genome Consortium 2012) was used to identify sites in the following classes: intergenic, intronic and codon positions one, two and three. CpG islands were identified using the CpGcluster Perl script (Hackenberg *et al.* 2006), and these regions were excluded from the above classes.

We next identified all fourfold degenerate (4D) sites in the genome as putative sites that experience little or no selection. To be defined as 4D, a third codon position had to meet two requirements: The first and second positions in the codon had to be invariable across all 80 samples, and there had to be no change to the encoded amino acid with any of the four possible bases in the third position (using the standard genetic code). This conservative approach identified 1,868,350 4D sites (12% of coding sequence, 0.7% of the whole genome).

Codon usage was assessed using the method of Wright (Wright 1990), which compares the effective number of codons, N_c, against the GC content at third codon positions for each gene (Fig. S3). Both values were estimated using the program codonW (http://codonw.sourceforge.net). In order to account for the effects of CUB in our downstream analyses, we defined a set of genes with minimal CUB, which deviated less than 5% from a neutral expectation for codon usage (Fig. S3), which constituted 7721 genes (61% of the total gene set).

Assigning scaffolds to chromosomes

Scaffolds were assigned to chromosomes based on the RAD-seq linkage map of the *Heliconius melpomene* genome v1.1 (The Heliconius Genome Consortium 2012), which has ~83% of the genome assigned to chromosomes. As many Z-linked scaffolds were not identified in that study, a specific assignment of scaffolds to the Z chromosome based on comparative read depth in males

and females was performed as part of the study of Martin et al. (Martin *et al.* 2013) and is provided as supplementary material therein. This procedure identified a large number of additional Z-linked scaffolds, and also several miss-assembled scaffolds that were Z/autosome chimeras. Using the most likely breakpoints, Z-linked regions were removed from autosomal scaffolds as were autosomal regions from Z-linked scaffolds. For plots across chromosomes, scaffolds were arranged according to the *Heliconius melpomene* genome (v1.1) linkage map (The Heliconius Genome Consortium 2012), after correcting for the several Z/autosome chimeric scaffolds, as well as manually correcting the orientation of several scaffold pairs based on mate-paired sequences.

Estimating the extent of linkage disequilibrium

Linkage disequilibrium (LD) was estimated for Eastern and Western populations using all pairs of biallelic sites with high-quality genotype calls for at least 50 of the 58 *H. melpomene* samples and a minor allele count of at least 5. We estimated r^2 using the maximum likelihood estimator of Clayton and Leung (Clayton and Leung 2007), implemented in the R package *snpstats*, which does not require phased haplotypes. To investigate how LD breaks down with distance, r^2 values were binned according to distance in logarithmically increasing bin sizes, which accounts for small numbers of SNP pairs at large distances. The top 100 longest scaffolds were analysed, and only SNP pairs on the same scaffold were considered. To obtain an estimate of background LD between unlinked sites, subsets of 500 SNPs were randomly selected and r^2 was estimated for all pairs for which the two SNPs were on separate chromosomes. This procedure was repeated 100 times and the mean was taken.

Analysis of population structure

We ran STRUCTURE for the 58 wild *H. melpomene* samples, together with the four *H. cydno* and four *H. timareta* samples, and tested k values from 3 to 8. Admixture between clusters was allowed, and correlated allele frequencies between clusters were assumed. Each run consisted of a burn-in of 10,000 iterations followed by another 10,000 generations, and five runs were performed for each k value to check consistency. To minimise the effects of direct selection, only four-fold degenerate sites were considered. Sites were required to be genotyped in at least 50 of the 66 analysed samples, and to have a minor allele count of at least two.

The second method for population structure analysis was a Principle Components Analysis (PCA), performed using Eigenstrat (Price *et al.* 2006). PCA is a method to simplify correlated multidimensional data, and can therefore reduce DNA sequence data for many SNPs to a small number of principal components that capture most of the information about the relationships among the sequences. The same filtered SNP dataset used for STRUCTURE was used, except that the *H. cydno* and *H. timareta* samples were excluded. Eigenstrat incorporates chromosomal information, but assumes that data is from humans. The 20 *Heliconius* autosomes were therefore labelled as human chromosomes 1-20, and the *Heliconius* Z chromosome as human chromosome 23 (X).

Site frequency spectra

Unfolded site frequency spectra were generated by counting the number of derived alleles at each site in each *H. melpomene* population. Sites were polarised using the four silvaniform outgroup samples, and only biallelic sites where the silvaniforms were fixed for one of the two alleles segregating in *H. melpomene* were considered, with said allele designated as ancestral. Site frequency spectra can only be compared between samples of the same size. The smallest population sample was for the Colombian population, with just five diploid individuals. To compare all of the populations, frequency spectra were estimated by sampling five individuals to represent each site in each population. Sites with fewer than five samples genotyped were ignored. To compare just the Eastern and Western populations, which had more samples, 20 individuals were sampled per site. Because there were females (heterogametic, ZW) in the dataset, only autosomes were considered so that all individuals were diploid at all sites.

Inference of ancestral population size using PSMC

Beginning with binary sequence alignment map (bam) files, we followed the authors' suggestions for genotyping using Samtools (Li *et al.* 2009), which involved a quality cut-off of 20 and depth cut-offs of 1 third (minimum) and two thirds (maximum) of mean depth for each sample. PSMC was run with 25 iterations, with 29 interval parameters spread over 58 time intervals (with the command flag -p “28*2+3+5”). A generation time of 0.25 years and a mutation rate of $2x10^{-9}$ was used. A number of different block sizes were considered, but reported results used the default of 100 bp.

Window-based population parameters: Various population parameters were calculated for non-overlapping 100 kb windows across the genome. Only windows with a sufficient number of sites genotyped in at least 50% of samples were considered. This number depended on the site classes being considered; all site classes: 10 000; intergenic: 6 000; intronic: 5 000; codon positions 1, 2 or 3: 500; 4D sites: 250; 4D sites in low-CUB genes: 150. We used 100 kb windows because linkage disequilibrium (LD) tends to break down almost completely within 10 kb, and reaches background levels within 100 kb (Fig. S1), meaning that measures from adjacent windows would be largely free of linkage effects.

Nucleotide diversity (π) and absolute divergence (d_{XY}), were calculated as the average proportion of differences between all pairs of sequences, either within a sample (π) or between two samples (d_{XY}). We used custom-written functions that ignored missing data in a pair-wise manner to maximise the amount of data being considered. Tajima's D (Tajima 1989) and F_{ST} (as in equation 9 of Hudson et al. [1992]) were calculated using the EggLib Python module (De Mita and Siol 2012).

Poor mapping and genotyping in highly variable regions could lead to underestimates of diversity and divergence. If the most variable parts of the genome are clustered, then we might expect windows with large amounts of missing data to also have higher levels of diversity and divergence at the sites that were genotyped. We therefore tested whether observed diversity and divergence estimates were correlated with the proportion of missing data per window (measured as the proportion of sites genotyped in fewer than 50% of individuals). We tested both third codon positions and intergenic sites.

Estimating the genome-wide rate of adaptive substitution

We estimated the genome-wide rate of adaptive substitution (α) using Messer and Petrov's asymptotic method (Messer and Petrov 2013). Briefly, α estimates the excess of between-species divergence at non-synonymous sites relative to synonymous sites, by comparison with the ratio of within-species polymorphism at these two site classes (McDonald and Kreitman 1991; Fay *et al.* 2001). The asymptotic method accounts for a number of potential confounding factors, including linked selection and segregating deleterious variation, by considering polymorphisms at each possible derived allele frequency separately, and fitting an exponential curve to the resulting

estimates, the asymptote of which should approximate the true α (Messer and Petrov 2013).

We used the Western population to measure polymorphism, because of its more stable population history. Due to the sensitivity of this method to small perturbations in the site frequency spectrum, we first filtered out all SNPs that failed a test for Hardy-Weinberg equilibrium (Wigginton *et al.* 2005) with P<0.05. Unfolded site frequency spectra were generated for synonymous and non-synonymous SNPs in all protein-coding genes, polarized using the four *H. erato* samples. Each SNP was randomly down-sampled to 16 individuals and, for Z linked genes only males (diploid, ZZ) were considered. By using *H. erato* as the outgroup, the class of sites where the Western population was fixed for the derived allele provided our measure of divergence, and all other classes where the derived allele was not fixed provided the measures of polymorphism. Following Messer and Petrov, we fitted an asymptotic curve to the estimated α values for all derived allele frequencies greater than 0.1.

The rate of synonymous and non-synonymous substitutions

We used Codeml from the PAML package (v. 4.8) (Yang 1997, 2007) to estimate ω, the ratio of synonymous substitutions per synonymous site to non-synonymous substitutions per non-synonymous site (d_N/d_S). We used a five-species input tree, including one species from each of the five major clades, with the topology (((*melpomene*,*hecale*),*wallacei*),*hecuba*,*erato*), following our maximum-likelihood phylogeny (Fig. 1B) and the genus-wide multilocus tree of Kozak et al. (Kozak *et al.* 2015). Each species was represented by a single high-coverage sample: *H. melpomene* (melP.HGC1), *H. hecale* (hec.JM273), *H. wallacei* (wal.CJ8687), *H. hecuba* (hecu.CJ8550) and *H. erato* (era.CJ2980). Because Codeml does not accommodate heterozygous genotypes, we selected a single allele to represent each species at each heterozygous site. To minimise the contribution of intra-specific variation to measures of inter-specific divergence, we selected the allele that was most commonly represented in the other four species at that site, and therefore likely to be ancestral.

For each 100 kb window, the coding sequences for all genes in the window were concatenated for analysis in Codeml, and only windows with genotype calls for all five species at at least 500 coding sites were included. We ran Codeml for each window via EggLib (De Mita and Siol 2012), using the “branch model” (Yang and Nielsen 1998; Yang 1998), which allows for different branches

of the tree to have different values of ω (d_N/d_S), the ratio of non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site. As we were most interested in substitutions along the lineage leading to *H. melpomene*, we constrained other branches to a single ω value, with the terminal branch leading to *H. melpomene* (after the split from *H. hecale*) allowed a distinct ω value.

Rates of adaptive substitution for individual genes

We also estimated *a*, the number of adaptive non-synonymous substitutions, for individual protein-coding genes, using the maximum likelihood method implemented in the program Mktest (v2) (Welch 2006). Default parameters were used except that *f*, which determines the strength of purifying selection, and *a* was allowed to vary among loci. Polymorphism was calculated for the Western population, and divergence was calculated between the Western samples and the four *H. erato* samples. To minimize missing data, three samples with fewer than 82% of coding sites genotyped (compared to >90% in all other samples) were excluded. Missing genotypes in the remaining samples were substituted with the major allele. To account for the presence of mildly-deleterious alleles contributing to non-synonymous polymorphism, we imposed a minor allele frequency cut-off, implemented by converting minor alleles occurring below this frequency to the major allele. We tested a range of cutoffs and selected a value of 0.24, which appeared to eliminate most segregating deleterious variation (data not shown). Nevertheless, it has been shown that imposing an arbitrary cut-off may still fail to exclude all segregating deleterious variants (Charlesworth and Eyre-Walker 2008; Messer and Petrov 2013). However, we are primarily interested in the variation among genes and not the absolute value of *a*.

Local recombination rate estimation

We estimated local recombination rates using the *Heliconius melpomene* genome (v1.1) linkage map (The Heliconius Genome Consortium 2012). We first corrected for the several Z/autosome chimeric scaffolds as mentioned above, and manually corrected the orientation of several scaffold pairs based on mate-paired sequences. We used local regression (loess) to fit physical distance to marker distance for each chromosome, with a span equivalent to 5 Mb. The recombination rate for each window was then estimated by taking the gradient of the smoothed curve (i.e. bp per CM) between the start and end points of each window. Corbett-Detig et al. (Corbett-Detig *et al.*

2015) recently used a similar approach to estimate recombination rates from the same map, but using different smoothing approaches. We tested their estimated recombination rates in our multiple regression model, and the result was nearly identical (data not shown). We therefore only report results from analyses using our estimates.

Multiple Regression

We used multiple linear regression to model nucleotide diversity at 4D sites (π_{4D}) in 100 kb windows (calculated for each population separately and then averaged).The following explanatory variables were included: local gene density (the proportion of coding sequence per window), local recombination rate ($\hat{r}$) across the window, the number of recent non-synonymous substitutions (D_n) in the *H. melpomene* lineage per window, the rate of synonymous substitutions per synonymous site (d_S) and GC-content at third codon positions.

To reduce noise caused by a lack of data in some windows, only windows with at least 250 4D sites genotyped in at least 50% of individuals were considered. Windows overlapping scaffold edges by more than 50% were discarded. Only windows on autosomal scaffolds were considered. This left 1467 windows for fitting the model. Because these still varied in the amount of available data, residuals were weighted according to the number of analysed 4D sites with genotype calls for at least 50% of individuals.

All variables included in the multiple regression were transformed to reduce skewness in their distributions. The response variable, π_{4D}, was square-root transformed. Explanatory variables d_S and D_n and a were square-root transformed, GC-content was $\log_{10}$ transformed and gene density was logit transformed. To allow for comparison of effects, all explanatory variables were then also Z-transformed. Residual analysis of the model revealed no indication that model assumptions were violated.

To further investigate the interrelationships between the explanatory variables, we used principal component regression (PCR) (Drummond *et al.* 2006; Mugal *et al.* 2013). This approaches can help to tease apart the effects of the various explanatory variables by summarizing the explanatory variables into orthogonal components, thereby accounting for multi-collinearity. Regression analyses were performed with the R version 3.0.3 (https://www.R-project.org) using the *pls* package

(Mevik BH, Wehrens R 2013).

To assess the robustness of our findings, we investigated the influence of various modifications to the model. Firstly, because selection for codon usage can affect diversity at 4D sites, we also tested models in which the response variable, π_{4D}, was calculated only using genes showing minimal evidence of codon usage bias. Chromosome ends may have reduced recombination rates, and could strongly influence the results of our analysis, so we also tested the model excluding windows in the outer 5% of each chromosome. An additional concern was that missing data in highly divergent genes might bias the detection of adaptive non-synonymous substitutions downwards. To account for this bias we tested a model using D_n and d_S calculated using only *H. melpomene* and it's closer relatives the silvaniforms, excluding the more distantly related *H. Wallacei*, *H. hecuba* and *H. erato*. This would reduce the effect of missing data, but could add noise due to the smaller time-scale being assessed. Lastly, as mentioned above, we tested a model with the estimated gene-by-gene number of adaptive non-synonymous substitutions (*a*, summed for all genes per window) used in place of D_n to account for the effects of hitchhiking. Due to the large amounts of noise in these genic *a* estimates, the distribution had long tails, and windows below the 10th and above the 90th percentiles for *a* were excluded from this model.

Testing for reduced diversity around non-synonymous substitutions

In addition to the multiple regression model, we also tested for a directly observable reduction in diversity around non-synonymous substitutions, after accounting for mutation rate variation. Recent substitutions unique to the *H. melpomene* lineage were identified as sites where at least two outgroup lineages carried the same (ancestral) allele, but where *H. melpomene* was fixed for a derived allele. Scaled diversity at 4D sites (π/d_{XY}) was calculated as a function of distance from the nearest substitution. To this end, each 4D site in the genome was binned according to its distance from the nearest non-synonymous substitution, in bins of 50 bp, up to a distance of 50 kb. Nucleotide diversity (π) in the Eastern and Western populations of *H. melpomene*, and absolute divergence (d_{XY}) between these two populations and *H. erato*, were calculated for each 4D site and averaged in each bin. Because fixed substitutions may be more common in low-diversity regions, we might expect reduced diversity near the site of substitutions regardless of whether hitchhiking has occurred. To account for this potential bias, we compared scaled diversity binned by distance

from non-synonymous substitutions to that binned by distance from synonymous substitution, using a bootstrapping procedure similar to that of Sattath et al. (2011). In each bootstrap replicate, a subset of synonymous substitutions, equal in size to the set of non-synonymous substitutions, was sampled with replacement, and scaled diversity was calculated in bins according to their distance from the nearest substitution in this subset. This procedure was repeated 100 times so that a confidence interval could be calculated.

Scanning for selective sweeps with SweeD

SweeD calculates a composite likelihood ratio (CLR) by comparing the site frequency spectra for individual blocks of sequence to that for the entire region. Therefore, sequences for whole chromosomes were produced by concatenating all scaffolds that mapped to each chromosome in the inferred order as described above. Although in many cases several scaffolds map to the same location, and their orientations are often unknown, this should not dramatically affect the method, as each window is considered independently and only a small subset of windows will cross scaffold boundaries. SweeD was run for each chromosome with a grid size of 1000 blocks. Sites were polarised (i.e. unfolded), where possible, by identifying the ancestral state based on the four silvaniform outgroups. If the outgroups were not fixed for a single allele, the site was designated as folded.

To select a threshold CLR value for identifying blocks with a significantly skewed frequency spectrum, we simulated sequence data for analysis in SweeD. For both the Eastern and Western populations, 100 simulated datasets, of 1 Mb each were simulated using ms (Hudson 2002). Theta (θ) values of 1.7% and 1.8% were used for the two respective populations, and a population recombination rate of 0.1% was used for both. We tested simulations based on the inferred population size histories from the PSMC analysis (see above) and with constant population size. The constant population size runs gave higher, more conservative threshold CLR values (data not shown), as was found by Nielsen et al. (Nielsen *et al.* 2005). The 99% quantile for the two populations were similar, so we used the higher value, 34, as the CLR cut-off for both analyses.

REFERENCES

Charlesworth J., Eyre-Walker A., 2008 The McDonald-Kreitman test and slightly deleterious mutations. Mol. Biol. Evol. **25**: 1007–15.

Clayton D., Leung H.-T., 2007 An R package for analysis of whole-genome association studies. Hum. Hered. **64**: 45–51.

Corbett-Detig R. B., Hartl D. L., Sackton T. B., 2015 Natural Selection Constrains Neutral Diversity across A Wide Range of Species. PLOS Biol. **13**: e1002112.

DePristo M. a, Banks E., Poplin R., Garimella K. V, Maguire J. R., Hartl C., Philippakis A. a, Angel G. del, Rivas M. a, Hanna M., McKenna A., Fennell T. J., Kernytsky A. M., Sivachenko A. Y., Cibulskis K., Gabriel S. B., Altshuler D., Daly M. J., 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. **43**: 491–8.

Drummond D. A., Raval A., Wilke C. O., 2006 A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. **23**: 327–37.

Fay J. C., Wyckoff G. J., Wu C.-I., 2001 Positive and Negative Selection on the Human Genome. Genetics **158**: 1227–1234.

Hackenberg M., Previti C., Luque-Escamilla P. L., Carpena P., Martínez-Aroza J., Oliver J. L., 2006 CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinformatics **7**: 446.

Huang W., Li L., Myers J. R., Marth G. T., 2012 ART: a next-generation sequencing read simulator. Bioinformatics **28**: 593–594.

Hudson R. R., Boos D., Kaplan N., 1992 A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9**: 138–51.

Hudson R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**: 337–338.

Kozak K. M., Wahlberg N., Neild A., Dasmahapatra K. K., Mallet J., Jiggins C. D., 2015 Multilocus Species Trees Show the Recent Adaptive Radiation of the Mimetic. Syst. Biol.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–9.

Lunter G., Goodson M., 2011 Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. **21**: 936–9.

Martin S. H., Dasmahapatra K. K., Nadeau N. J., Salazar C., Walters J. R., Simpson F., Blaxter M., Manica A., Mallet J., Jiggins C. D., 2013 Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. **23**: 1817–1828.

McDonald J. H., Kreitman M., 1991 Adaptive Protein Evolution at the Adh Locus in Drosophila. Nature **351**: 652–654.

Messer P. W., Petrov D. a., 2013 Frequent adaptation and the McDonald-Kreitman test. Proc. Natl. Acad. Sci. **110**: 8615–8620.

Mevik BH, Wehrens R L. K. H., 2013 pls: Principal component and partial least squares regression.

Mita S. De, Siol M., 2012 EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genet. **13**: 27.

Mugal C. F., Nabholz B., Ellegren H., 2013 Genome-wide analysis in chicken reveals that local levels of genetic diversity are mainly governed by the rate of recombination. BMC Genomics **14**: 86.

Nielsen R., Williamson S., Kim Y., Hubisz M. J., Clark A. G., Bustamante C., 2005 Genomic scans for selective sweeps using SNP data. Genome Res. **15**: 1566–75.

Price A. L., Patterson N. J., Plenge R. M., Weinblatt M. E., Shadick N. A., Reich D., 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. **38**: 904–909.

Rambaut A., Grass N. C., 1997 Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics **13**: 235–238.

Sattath S., Elyashiv E., Kolodny O., Rinott Y., Sella G., 2011 Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in Drosophila simulans. PLoS Genet. **7**: e1001302.

Tajima F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–595.

The Heliconius Genome Consortium 1, 2012 Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature **487**: 94–8.

Welch J. J., 2006 Estimating the genomewide rate of adaptive protein evolution in Drosophila. Genetics **173**: 821–837.

Wigginton J. E., Cutler D. J., Abecasis G. R., 2005 A note on exact tests of Hardy-Weinberg equilibrium. Am. J. Hum. Genet. **76**: 887–893.

Wright F., 1990 The “effective number of codons” used in a gene. Gene **87**: 23–29.

Yang Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**: 555–556.

Yang Z., Nielsen R., 1998 Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. **46**: 409–18.

Yang Z., 1998 Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15**: 568–73.

Yang Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**: 1586–91.